Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information

Bilel Moulahi

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*
Cotutelle internationale *Université de Tunis El Manar*

Présentée et soutenue le *11/12/2015* **par :**

BILEL MOULAHI

Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information

---

JURY

PATRICE BELLOT	Professeur, Université Aix-Marseille	Rapporteur
CHIRAZ LATIRI	MCF/HDR, Université de la Manouba	Rapporteur
MOHAND BOUGHANEM	Professeur, Université de Toulouse 3	Examinateur
GABRIELLA PASI	Professeur, Università di Milano Bicocca	Examinatrice
LYNDA TAMINE	Professeur, Université de Toulouse 3	Directrice
SADOK BEN YAHIA	Professeur, Université de Tunis El Manar	Directeur

---

École doctorale et spécialité :
MITT : Image, Information, Hypermedia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)
Directeur(s) de Thèse :
Lynda Tamine et *Sadok Ben Yahia*
Rapporteurs :
Patrice Bellot et *Chiraz Latiri*

Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information

Bilel MOULAHI

2 janvier 2016

Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information
Manuscrit soumis pour le diplôme de Docteur de Philosophie
Doctorat soutenu le 11-12-2015
Doctorant : Bilel Moulahi
Directeurs de thèse : Lynda Tamine et Sadok Ben Yahia

Commentaires, corrections, et autres remarques sont les bienvenus :
moulahi@irit.fr, tamine@irit.fr, sadok.benyahia@fst.rnu.tn

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS,
Université Toulouse 3 Paul Sabatier,
118 route de Narbonne,
F-31062 Toulouse CEDEX 9

Dédicace

À mes parents,
À mes trois chers frères, Ammar, Taher et Mohsen,
À ma chère Mouna, à Hinda, Aida et Salwa ,
À ma chère Eya,
À mes petits : Eya, Hideya, Ahmed, Abouda, Baha, Taha et Hamma,
À la mémoire de mon cher ami Hosni,
À tous mes amis,
À la mémoire de khalti Salwa.

Remerciements

Par ces quelques lignes, je souhaite remercier toutes les personnes qui ont contribué de près ou de loin à l'aboutissement de cette thèse.

Je voudrais avant tout adresser mes plus chaleureux remerciements à mes deux directeurs de thèse Lynda Tamine et Sadok Ben Yahia.

Je tiens à exprimer ma très profonde gratitude à Lynda pour la confiance qu'elle m'a accordée en acceptant de diriger ma thèse, alors que j'ignorais tout de la Recherche d'Information. Je la remercie très sincèrement pour sa disponibilité, son soutien et tous ses précieux conseils au cours de la thèse.Je suis fier de l'avoir eu comme *mentor* et d'avoir appris à ses côtés la rigueur scientifique et la pédagogie pour présenter et rédiger les travaux pendant ces années de thèse. Merci spécialement pour cette dernière année et toute l'aide qu'elle m'a prodiguée pour la préparation des dossiers de candidatures, et en parallèle les relectures acharnées de ce manuscrit et son investissement pour la réalisation de nos différents travaux et publications ! Merci aussi pour toutes les réunions, parfois les soirs sur *Skype*, et même quand je rentrais en Tunisie. Je la remercie également de m'avoir accordé sa confiance lors des enseignements que j'ai réalisés. Je ne la remercierai jamais assez et je lui serai toujours reconnaissant.

Je souhaite également exprimer mes sincères remerciements et reconnaissance à Sadok, qui m'encadre depuis 2009 (7 années !), et avec qui j'ai effectué mes tous premiers stages de recherche en Maîtrise et en Master. Je le remercie pour son souci constant de l'avancement de ma thèse ainsi que son suivi continu de mon travail. Ses conseils, ses encouragements ainsi que la confiance qu'il m'a toujours témoigné m'ont été d'un grand apport tout au long de mes travaux. Monsieur, je n'oublierai jamais que c'était grâce à toi que je fais de la recherche aujourd'hui.

Je remercie tous les membres du jury d'avoir accepté de participer à l'évaluation de ce travail. Merci à Mme Chiraz Latiri et M. Patrice Bellot d'avoir rapporté mon mémoire de thèse, ainsi que pour leurs remarques encourageantes et constructives qu'ils m'ont données. Je remercie également Mme Gabriella Pasi et M. Mohand Boughanem qui m'ont fait honneur en acceptant de faire partie de mon jury de thèse.

Je remercie l'équipe SIG et la direction du laboratoire IRIT de m'avoir accueilli chaleureusement durant cette thèse. Des remerciements tout particuliers à Gilles et Yoann de

m'avoir accordé leur confiance en me laissant participer à leurs enseignements. Je ne peux pas oublier Chantal Morand, Jean-Pierre Baritaud et tout le personnel de l'IRIT pour leur aide durant ces années au laboratoire. Merci également à l'ensemble des doctorants : *Rafiiik*, pour les nombreuses soirées et nuits blanches qu'on avait passées au labo, Baptiste, Thomas, Gia et Arlind pour leur bonne humeur et les longues discussions durant les pauses cafés sur l'histoire, le sport et la politique. C'est un véritable plaisir de partager le bureau avec vous ! Je n'oublie pas les anciens membres de l'équipe qui étaient là quand je suis arrivé. Un grand merci à *Lauuure* pour ses conseils et encouragements durant toute ma thèse, Amjed pour son aide, Firas pour les matches de foot qu'il a perdu, Adrian, Bastien... Je pense aussi à Mohamed, qui m'a beaucoup aidé, Ismail, Eya, Manel, Imen, Hamid, Jonathan, sans oublier Ameni et Mariem qui viennent d'arriver. J'associe à ces remerciements l'ensemble de mes collègues et amis de la FST : Nidhaal, Imen, Hmida, Malek, Aymen, Rami, Khawla, Nejeh, Sawssen, Zarrouk et les autres... Je pense aussi à Chiraz Trabelsi, avec qui j'ai collaboré durant mon stage de Master.

I would like to thank Michael Gertz for the opportunity to visit the University of Heidelberg in Germany during a two-month internship. I would especially like to express my gratitude to my friends there : Jannik, Hamed, Katarina, Florian, Christian, Ayser... Hope to see you somewhere in this little world.

Je remercie aussi mes chers amis qui ne cessent de m'encourager par téléphone, mail et sur Facebook : Nader, *rafi9i* depuis 10 ans, le cher Mohamed, Khalil Khalili, Mahdouch, Omar, sans oublier Mansour le joyeux, mon fidèle ami Ramzi Roger, Mohamed Mkachakh, Wajdi, Karim... Un coucou spécial à Lazz pour tout ce qu'il a fait pour moi pendant mes premiers séjours en France.

Les derniers mais non les moindres, je tiens à remercier chaleureusement mes parents, mes trois grands frères Ammar, Taher et Mohsen, ma petite sœur adorée, Mouna, pour leur soutien sans fin. Je leur dois beaucoup pour ce qu'ils ont fait pour moi, leur sacrifices, leur prières. Que vous trouviez ici l'expression de ma gratitude et ma reconnaissance. Je n'aurais jamais réussi à en arriver là sans votre amour et vos encouragements. Merci à mes belle-sœurs pour leur soutien et encouragements continus. A mes petits neveux et nièces, je prie le bon Dieu de me donner les moyens de toujours prendre soin de vous. Désolé d'être trop absent. Je vous aime. Je remercie également mes chers cousins, Radhouane, Moez, Ayoub, Seddik, Ali, Seif et tous les autres, mes frères et amis depuis mon enfance. Je n'oublie pas ma tante Mariem, Noura, Imen, Asma, Omayma et tous les autres.

Mention spéciale à ma jolie fiancée, Eya, qui a été là pour moi à chaque étape de cette expérience. Son amour et son soutien ont beaucoup contribué à mon succès. Merci *ayouta* de ta patience, de ton support et de pouvoir subir mon stress durant ces trois dernières années. Bientôt, tu seras à Toulouse ! ! ! c'est la première fois que je sens le goût du succès accompagné par un bonheur complet. Finally ! We're getting married ...

Abstract

The main research topic of this document revolve around the information retrieval (IR) field. Traditional IR models rank documents by computing single scores separately with respect to one single objective criterion. Recently, an increasing number of IR studies has triggered a resurgence of interest in redefining the algorithmic estimation of relevance, which implies a shift from topical to multidimensional relevance assessment. In our work, we specifically address the multidimensional relevance assessment and evaluation problems. To tackle this challenge, state-of-the-art approaches are often based on linear combination mechanisms. However, However, these methods rely on the unrealistic additivity hypothesis and independence of the relevance dimensions, which makes it unsuitable in many real situations where criteria are correlated. Other techniques from the machine learning area have also been proposed. The latter learn a model from example inputs and generalize it to combine the different criteria. Nonetheless, these methods tend to offer only limited insight on how to consider the importance and the interaction between the criteria. In addition to the parameters sensitivity used within these algorithms, it is quite difficult to understand why a criteria is more preferred over another one.

To address this problem, we proposed a model based on a multi-criteria aggregation operator that is able to overcome the problem of additivity. Our model is based on a fuzzy measure that offer semantic interpretations of the correlations and interactions between the criteria. We have adapted this model to the multidimensional relevance estimation in two scenarii : *(i)* a tweet search task and *(ii)* two personalized IR settings.

The second line of research focuses on the integration of the temporal factor in the aggregation process, in order to consider the changes of document collections over time. To do so, we have proposed a time-aware IR model for

combining the temporal relavance criterion with the topical relevance one. Then, we performed a time series analysis to identify the temporal query nature, and we proposed an evaluation framework within a time-aware IR setting.

Résumé

La problématique générale de notre travail s'inscrit dans le domaine scientifique de la recherche d'information (RI). Les modèles de RI classiques sont généralement basés sur une définition de la notion de pertinence qui est liée essentiellement à l'adéquation thématique entre le sujet de la requête et le sujet du document. Le concept de pertinence a été revisité selon différents niveaux intégrant ainsi différents facteurs liés à l'utilisateur et à son environnement dans une situation de RI. Dans ce travail, nous abordons spécifiquement le problème lié à la modélisation de la pertinence multidimensionnelle à travers la définition de nouveaux modèles d'agrégation des critères et leur évaluation dans des tâches de recherche de RI. Pour répondre à cette problématique, les travaux de l'état de l'art se basent principalement sur des combinaisons linéaires simples. Cependant, ces méthodes se reposent sur l'hypothèse non réaliste d'additivité ou d'indépendance des dimensions, ce qui rend le modèle non approprié dans plusieurs situations de recherche réelles dans lesquelles les critères étant corrélés ou présentant des interactions entre eux. D'autres techniques issues du domaine de l'apprentissage automatique ont été aussi proposées, permettant ainsi d'apprendre un modèle par l'exemple et de le généraliser dans l'ordonnancement et l'agrégation des critères. Toutefois, ces méthodes ont tendance à offrir un aperçu limité sur la façon de considérer l'importance et l'interaction entre les critères. En plus de la sensibilité des paramètres utilisés dans ces algorithmes, est très difficile de comprendre pourquoi un critère est préféré par rapport à un autre.
Pour répondre à cette première direction de recherche, nous avons proposé un modèle de combinaison de pertinence multicritères basé sur un opérateur d'agrégation qui permet de surmonter le problème d'additivité des fonctions de combinaison classiques. Notre modèle se base sur une mesure qui permet

de donner une idée plus claire sur les corrélations et interactions entre les critères. Nous avons ainsi adapté ce modèle pour deux scénarios de combinaison de pertinence multicritères : *(i)* un cadre de recherche d'information multicritères dans un contexte de recherche de tweets et *(ii)* deux cadres de recherche d'information personnalisée.
Le deuxième axe de recherche s'intéresse à l'intégration du facteur temporel dans le processus d'agrégation afin de tenir compte des changements occurrents sur les collection de documents au cours du temps. Pour ce faire, nous avons proposé donc un modèle d'agrégation sensible au temps pour combinant le facteur temporel avec le facteur de pertinence thématique. Dans cet objectif, nous avons effectué une analyse temporelle pour éliciter l'aspect temporel des requêtes, et nous avons proposé une évaluation de ce modèle dans une tâche de recherche sensible au temps.

Publications

Articles de revues internationales

Bilel Moulahi, Lynda Tamine, Sadok Ben Yahia. When time meets information retrieval : Past proposals, current plans and future trends. In *Journal of Information Science (JIS)*. 2015. Sage (à paraître).

Bilel Moulahi, Lynda Tamine, Sadok Ben Yahia. iAggregator : Multidimensional relevance aggregation based on a fuzzy operator. In *Journal of the Association for Information Science and Technology (JASIST)*. Vol. 64, N. 10, pages 2062-2083, 2014. Wiley.

Conférences internationales

Bilel Moulahi, Lynda Tamine and Sadok Ben Yahia. Leveraging Temporal Query-Term Dependency for Time-Aware Information Access (regular paper). In *IEEE/WIC/ACM International Conference on Web Intelligence (WI 2015)*. Singapour, December 6-9, 2015. IEEE (à paraître).

Bilel Moulahi, Lynda Tamine and Sadok Ben Yahia. Toward a Personalized Approach for Combining Document Relevance Estimates (regular paper). In *Proceedings of the 22nd Conference on User modelling, Adaptation and Personalization (UMAP 2014)*. Vol. 8538, Pages 158-170, Aalborg, Denmark, July 7-11, 2014. Springer.

Article de campagnes d'évaluation internationales

Bilel Moulahi, Jannik Strötgen, Michael Gertz and Lynda Tamine. HeidelToul : A Baseline Approach for Cross-document Event Ordering (regular paper). In *T9th International Workshop on Semantic Evaluation (SemEval'15) (together with NAACL-HLT'15)*. Denver, Colorado, June 4-5, 2015.

Rafik Abbes, Bilel Moulahi, Abdelhamid Chellal, Karen Pinel-Sauvagnat, Nathalie Hernandez, Mohand Boughanem, Lynda Tamine, and Sadok Ben Yahia . IRIT at TREC 2015 Temporal Summarization Track (regular paper). In *Text REtrieval*

Conference (TREC 2015). Gaithersburg, USA, 2015. NIST. (à paraître).

Abdelhamid Chellal, Lamjed Ben Jabeur, Laure Soulier, Bilel Moulahi, Thomas Palmer, Mohand Boughanem, Karen Pinel-Sauvagnat, Lynda Tamine, and Gilles Hubert. IRIT at TREC 2015 Microblog Track (regular paper). In *Text REtrieval Conference (TREC 2015)*. Gaithersburg, USA, 2015. NIST. (à paraître).

Bilel Moulahi, Lynda Tamine and Sadok Ben Yahia. IRIT at TREC 2014 Contextual Suggestion Track (regular paper). In *Text REtrieval Conference (TREC 2014)*. Gaithersburg, USA, November 18-21, 2014. NIST.

Conférences et Workshops nationaux

Bilel Moulahi, Lynda Tamine, Sadok Ben Yahia. Prise en compte des préférences des utilisateurs pour l'estimation de la pertinence multidimensionnelle d'un document (regular paper). In *INFormatique des ORganisation et Systèmes d'Information de Décision (INFORSID 2014)*. Pages 295-310, Lyon, France, 20-23 Mai, 2014.

Bilel Moulahi, Lynda Tamine, Sadok Ben Yahia. L'intégrale de Choquet discrète pour l'agrégation de pertinence multidimensionnelle (regular paper). In *Conférence francophone en Recherche d'Information et Applications (CORIA 2013)*. Pages 399-414, Neuchâtel, Suisse, 3-5 Avril, 2013.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

1.1 Les modèles de recherche d'information classique et estimation de la pertinence

La recherche d'information (RI) est un domaine de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information. Ce besoin est souvent formulé en langage naturel et exprimé sous forme d'une requête décrite par un ensemble de mot clés. L'essor du web et la popularisation d'internet ont propulsé l'information au premier plan et ont ainsi remis la RI face à de nouveaux défis, à savoir *(i)* retrouver une information pertinente dans une masse grandissante de documents qui change dans le temps et qui *(ii)* répond aux besoins spécifiques de l'utilisateur. En effet, la limite majeure des modèles classiques de RI réside en partie dans le fait qu'ils sont basés sur une approche généraliste, qui considère que le besoin en information est complètement représenté par sa requête et délivrant alors des résultats ne tenant compte que des critères de l'adéquation thématique entre les requêtes et les documents (Vickery, 1959; Cooper, 1971; Harter, 1992). De nombreux modèles, tels que le modèle vectoriel, le modèle probabiliste et le modèle de langue, traduisent la pertinence par une fonction d'appariement de distribution de termes dans les termes et documents. Ces travaux ont certes apporté des améliorations, mais les performances de ces systèmes dépendent de plusieurs facteurs problématiques. Dans la suite, nous présentons les principaux

facteurs qui ont entraîné l'émergence de la problématique liée au concept de pertinence.

1.2 De la pertinence thématique à la pertinence multidimensionnelle

Des réflexions ont été menées dans le but de mieux cerner la notion de pertinence du point de vue de l'utilisateur et d'identifier les différents facteurs ayant un impact sur cette notion. Les études menées dans ce sens (Borlund, 2003; Saracevic, 2007a) ont montré que la pertinence n'est pas une relation isolée entre un document et une requête, elle est définie en fonction du contexte dans lequel la recherche est effectuée. Plusieurs travaux ont suggéré d'investir des directions telles que la considération d'autres facteurs autres que le contenu des documents dans le processus de sélection de l'information pertinente.

Les études menées dans ce sens ont montré que la pertinence est définie selon différents niveaux intégrant différents facteurs liés à l'utilisateur, à la requête et à son environnement dans une situation de recherche d'information donnée (crédibilité et autorité des utilisateurs, diversité, fraîcheur des résultats de recherche, etc.).

C'est ainsi qu'une nouvelle direction de recherches basée sur la RI multi-critères est apparue. Cette direction de recherche prometteuse consiste à combiner des diverses sources d'évidence issues du contexte de l'utilisateur et de son environnement dans une même infrastructure afin de mieux caractériser les besoins en information et améliorer les résultats de recherche. C'est ainsi que plusieurs dimensions de pertinence ont vu leur intégration dans de nombreuses applications de RI :

- RI mobile (Göker et Myrhaug, 2008; Cong *et al.*, 2009; Bouidghaghen *et al.*, 2011b)
- RI sociale (Duan *et al.*, 2010; Nagmoti *et al.*, 2010; Ben Jabeur *et al.*, 2010; Metzler et Cai, 2011; Becker *et al.*, 2011; Ounis *et al.*, 2011; Chen *et al.*, 2012)
- RI personnalisée (Sieg *et al.*, 2007; Gauch *et al.*, 2003; Daoud *et al.*, 2010; Bouidghaghen *et al.*, 2011a; Daoud *et al.*, 2011; da Costa Pereira *et al.*, 2012)
- RI géographique (Mata et Claramunt, 2011; Kishida, 2010; Daoud et

Huang, 2013)

Chaque cadre de RI a sa spécificité en fonction des différents facteurs de pertinence qui entrent en jeu dans le processus d'ordonnancements de documents. En effet, ces critères ont de surcroît des poids d'importance variables selon la spécificité des tâches adressées, les caractéristiques intrinsèques des critères et le besoin en information des utilisateurs.

Le défi majeur qui se pose, est alors de trouver, en fonction du cadre, les méthodes appropriées pour pouvoir agréger les différentes dimensions ou critères afin d'aboutir à un seul score de pertinence des documents. Différents schémas basés sur la présence de critères, capturant des aspects potentiels sur les besoins des utilisateurs, ont été ainsi proposés.

1.3 Problématique

Dans cette thèse, nous nous intéressons à la définition de nouveaux modèles d'agrégation multicritères pour l'estimation de pertinence multidimensionnelle en RI. Nous considérons que chaque critère décline une facette de la pertinence avec un poids d'importance à prédire et que les scores de pertinences agrégés sont des fonctions à valeurs discrètes ou non discrètes, dépendantes ou indépendantes. La dépendance peut en outre s'avérer conflictuelle. L'obtention d'un score de pertinence global d'un document retourné, en réponse à une requête soumise par un utilisateur, consiste à agréger les scores de pertinence individuels par le biais d'un opérateur ou d'une fonction d'agrégation.

Pour répondre aux problèmes d'agrégation multicritères, les approches de l'état de l'art se sont généralement inspirées du contexte scientifique de la théorie des problèmes de prise de décision multicritères. Ces modèles se basent généralement sur la définition d'opérateurs d'agrégation s'appuyant sur des combinaisons linéaires ou non linéaires des scores de pertinence associés aux documents (Vogt et Cottrell, 1999; Larkey *et al.*, 2000; Si et Callan, 2002; Craswell *et al.*, 2005; Damak *et al.*, 2011; Wei *et al.*, 2011). Cependant, ces travaux applicatifs, en dépit de leur simplicité, ont généralement utilisé des fonctions de combinaison qui se basent sur l'hypothèse non réaliste d'additivité ou d'indépendance des dimensions. La propriété additive des mesures qu'exige ces méthodes les rend non appropriées dans plusieurs situations réelles dans lesquelles les critères étant (fortement) corrélés (ou présentant des interactions entre eux) (Saracevic, 2007b; Wolfe et Zhang,

2010; Carterette *et al.*, 2011; Eickhoff *et al.*, 2013a). En outre, ces limites ne permettent pas de représenter toutes les préférences des utilisateurs et conduisent ainsi à des résultats biaisés à cause des critères similaires qui peuvent affecter les autres critères et donner des scores d'agrégation surestimés (ou sous-estimés).

Partant de ce postulat, nous proposons de prendre en compte les relations pouvant exister entre les critères, et nous proposons un cadre d'agrégation qui permet d'identifier les préférences des utilisateurs. Nous nous intéressons plus spécifiquement à deux problématiques principales :

1. Comment modéliser le problème de combinaison de pertinence multidimensionnelle quand il s'agit de critères dépendants ?
   (a) Comment identifier les critères les plus importants dans l'évaluation globale de pertinence ?
   (b) Comment identifier les synergies ou dépendances entre ces critères ?
   (c) Comment inférer automatiquement les poids d'importance des dimensions de pertinence ?
   (d) Comment personnaliser l'opérateur d'agrégation pour considérer les préférences des utilisateurs ?
2. Comment intégrer le critère temporel dans le processus d'agrégation quand il s'agit des collections de documents qui évoluent dans le temps ?
   (a) Comment utiliser le type de requête pour identifier l'importance du critère temporel ?
   (b) Comment modéliser le critère de pertinence temporel avec d'autres dimensions dans un même schéma d'ordonnancement ?

1.4 Contributions

Afin de pallier le manque de flexibilité des modèles de combinaison classiques, nous proposons deux schémas d'agrégation dont l'un est dédié aux collections de documents statiques et l'autre est plus adapté aux flux de documents qui changent dans le temps. Pour chacune de ces approches, nous détaillons les contributions proposées ci-dessous.

1. *Une approche générique d'estimation de pertinence multidimensionnelle.* Le cadre général de nos travaux de thèse dans la première contribution concerne l'agrégation des dimensions de pertinence, qu'elles soient interdépendantes ou indépendantes. Dans un premier temps, nous proposons un modèle de combinaison de pertinence multicritères basé sur un opérateur flexible. Ce dernier est fondé sur les intégrales floues utilisées en aide à la décision multicritères (Grabisch, 1995). La principale originalité de cet opérateur réside dans sa capacité à modéliser des interactions entre les critères grâce à l'utilisation d'une mesure floue définie sur l'ensemble des critères. Ainsi, cette mesure permet de surmonter le problème d'additivité des fonctions de combinaison classiques, qui sont incapables de modéliser plusieurs situations du monde réel. Nous avons adapté ce modèle pour deux scénarios de combinaison de pertinence multicritères :

   (a) Une approche qui se base sur l'intégrale de Choquet discrète (Choquet, 1953; Grabisch, 1995; Grabisch et Labreuche, 2010), un opérateur qui a été largement exploité dans le domaine d'aide multicritères pour la prise de décision (Grabisch et Labreuche, 2010). Parmi nos motivations derrière l'adoption de ce type de méthodes, est d'essayer de résoudre en partie le problème de dépendance (ou corrélations) pouvant exister entre les critères de pertinence, comme déjà annoncé dans plusieurs travaux en RI (Carterette *et al.*, 2011; Eickhoff *et al.*, 2013b). Les principales contributions (Moulahi *et al.*, 2013, 2014d) dans cette première partie sont :

      i. L'adoption d'un opérateur mathématique qui d'un point de vue théorique, présente un certain nombre de propriétés qui semblent être très intéressantes sous l'angle de RI. Notre idée est de se baser sur la flexibilité et la capacité de cet opérateur dans la modélisation des dépendances pour éviter le biais introduit à cause des critères redondants ou complémentaires. L'intuition majeure consiste donc à définir des poids d'importance différents, non seulement sur les critères de pertinence individuels, mais aussi sur tous les sous ensembles de critères. Cette représentation robuste permet de faciliter l'interprétation des degrés d'importance des critères via l'indice d'interaction et l'indice de *Shapley* (Grabisch, 1996). Avec ces caractéristiques, notre méthode de combinaison pourrait être considérée comme une méta-classe qui permet de généraliser

la plupart des fonctions d'agrégation classiques.

ii. La proposition d'un algorithme supervisé pour l'apprentissage des poids d'importance des critères et sous ensembles de critères. Cet algorithme étant générique, ne dépend ni de la collection de données ni de la tâche de RI considérée. Ainsi, il permet de retourner des résultats qui sont facilement interprétables par des humains grâce aux deux indices d'interaction et de *Shapley*.

iii. Une évaluation approfondie du modèle d'estimation multi-critères dans une tâche de recherche de tweets (Ounis *et al.*, 2011, 2012). Nous avons appliqué le modèle dans une collection de test standard basée sur les tâches Microblog de TREC 2011 et 2012. Nous nous sommes basés sur des dimensions de pertinence déjà exploitées dans des travaux de RI sociale sur Twitter (Duan *et al.*, 2010; Nagmoti *et al.*, 2010).

Dans un second temps, nous présentons une approche d'agrégation personnalisée basée sur l'adaptation de la mesure floue sous-jacente à l'opérateur de Choquet Moulahi *et al.* (2014b,a,c). Cette partie comprend deux points clés :

i. Une agrégation pondérée par les préférences des utilisateurs. A travers la mesure floue, nous avons obtenu un schéma de pondération facilement personnalisable qui est à la base de la quantification de l'importance estimée de chaque dimension pour chaque utilisateur ainsi que leur degré d'interactivité ou d'interdépendance. Les degrés d'importance des critères sont estimés selon le même algorithme d'apprentissage déjà énoncé, en inférant les mesures optimales pour chaque utilisateur.

ii. Une évaluation de l'opérateur d'agrégation personnalisé dans deux contextes de RI différents, dont l'un en se basant sur un scénario de RI dans les folksonomies et l'autre en utilisant un contexte de RI contextuelle. Dans ces deux derniers scénarios, nous exploitons respectivement une collection de signets (*bookmark*) collectés à partir d'un système d'annotation sociale ainsi que la collection de test standard fournie par la tâche *Contextual Suggestion* de TREC 2014 (Dean-Hall *et al.*, 2013). Pour ces deux cadres de RI, nous montrons l'impact de la prise en compte des dépendances entre les critères de

pertinence ainsi que l'impact de leur personnalisation sur les performances de recherche.

2. *Une approche d'agrégation sensible au temps.* Le deuxième axe de recherche auquel nous nous sommes intéressés dans cette thèse est l'intégration du critère temporel dans le processus d'agrégation pour tenir compte des changements occurrants dans les collections de documents au cours du temps (Moulahi *et al.*, 2015a,c). Les principales contributions présentées pour répondre à cette problématique incluent :

   (a) Une approche d'estimation de pertinence multidimensionnelle sensible au temps basée sur l'injection du critère temporel au sein d'un modèle d'agrégation d'ordonnancements. Cette approche permet d'éliciter l'aspect temporel des requêtes en se basant sur les séries chronologiques (Moulahi *et al.*, 2015b). Ceci permet d'identifier les périodes auxquelles fait référence une requête et donc de favoriser les documents appartenant à cet intervalle de temps.

   (b) Une analyse temporelle des collections de test standards pour valider l'hypothèse de corrélation temporelle entre les termes d'une même requête. Nous avons ensuite évalué empiriquement notre approche sensible au temps sur les corpus fournis par les tâches Temporal Summarization de TREC 2013 et 2014.

1.5 Organisation de la thèse

Cette thèse est constituée d'un chapitre introductif ainsi que de trois principales parties, dont la première présente la synthèse des travaux de l'état de l'art, la seconde partie détaille nos principales contributions et la dernière conclut le manuscrit et discute des perspectives de recherche. Nous présentons le contenu ci-après.

Le chapitre 1 introduit la thèse. Il présente le contexte, les problématiques de recherche abordées et les contributions issues de nos travaux.

La première partie de cette thèse, intitulée *Synthèse des travaux de l'état de l'art* présente le contexte de nos travaux. Compte tenu du cadre de notre thèse, nous avons axé l'état de l'art sur les travaux de combinaison multicritères ainsi que sur les méthodes d'apprentissage d'ordonnancements et d'agrégation sensibles au temps. Cette partie englobe trois chapitres :

– Le **deuxième chapitre** intitulé "*Concepts de base de la RI classique*" introduit les principaux concepts de base du domaine de RI. Il présente les principaux modèles de RI ainsi que les mesures d'évaluation utilisées pour le test des différentes approches proposées dans ce cadre. Une formalisation de tous les modèles et métriques est ainsi proposée. Ce chapitre se termine par l'étude de l'émergence de la notion de pertinence multidimensionnelle. Il montre aussi l'orientation des travaux vers la RI multicritères et aborde les problématiques majeures et les verrous scientifiques.
– Le **troisième chapitre** intitulé "*Approches multicritères pour l'estimation de pertinence des documents en RI*" présente une revue critique de l'état de l'art et des différentes approches proposées pour l'agrégation multicritères. Il montre ensuite le principe d'agrégation de pertinence multidimensionnelle en RI. Une formalisation des approches d'agrégation multicritères issues des problèmes de prise de décision et des méthodes d'agrégation et apprentissage d'ordonnancements sont également présentées.
– Le **quatrième chapitre** intitulé "*Recherche d'information temporelle et pertinence : synthèse des travaux de l'art*" est dédié aux travaux de la littérature exploitant le temps dans le cadre des tâches de RI. Il présente une définition générale des concepts des critères fraîcheur d'information et récence et propose un schéma général pour catégoriser les travaux de l'état de l'art suivant la manière avec laquelle l'information temporelle a été exploitée. Un aperçu sur les collections de test standards existants ainsi que les cadres d'évaluation des systèmes de RI sensibles au temps qui pourraient être exploités est également présenté dans ce chapitre.

La deuxième partie de cette thèse, intitulée *Contribution à la définition et l'évaluation de modèles d'agrégation de pertinence multidimensionnelle en RI*, présente nos contributions relatives à l'agrégation de pertinence multidimensionnelle. Elle englobe deux chapitres présentant deux modèles différents pour la combinaison multicritères :

– Le **cinquième chapitre** intitulé "*Méthode d'agrégation de pertinence multidimensionnelle : proposition et évaluation dans des tâches de RI sociales et personnalisées* " est dédié à la présentation de la première contribution de la thèse. Ce chapitre dresse la problématique et quelques motivations puis présente une formalisation du problème d'agrégation de pertinence multidimensionnelle. Ensuite, il présente le modèle d'agrégation basé sur l'intégrale de Choquet discrète, et illustre la particularité qu'offre cette méthode pour éliciter les degrés d'importance des critères et identifier les dépendances pouvant exister entre eux.

La deuxième partie de ce chapitre intitulé "*Méthode d'agrégation personnalisée de pertinence : évaluation dans une tâche de recherche d'information personnalisée*" aborde le problème de personnalisation des préférences utilisateurs dans l'agrégation multicritères. Il décrit une méthode basée sur l'intégrale de Choquet permettant de personnaliser les poids d'importance des critères grâce à la flexibilité du concept de mesure floue. Ce chapitre présente enfin deux cadres d'évaluation dont l'une est dans un cadre de recherche de tweets et au sein d'une collection de test standard fournie par la tâche Microblog de TREC. Tandis que l'autre se situe au niveau des tâches TREC dédiées à la RI personnalisée en l'occurrence TREC Contextual Suggestion. L'approche est également évaluée dans une tâche de recherche personnalisée dans les folksonomies. Nous dressons le cadre expérimental puis les résultats de l'application des deux méthodes proposées. Ce chapitre est clôturé par une discussion des résultats dans les deux cadres d'évaluation ainsi qu'une étude de l'importance des différents critères utilisés.

– Le **sixième chapitre** intitulé "*Vers une approche d'agrégation guidée par la requête : évaluation dans le cadre d'une tâche de RI sensible au temps*" présente la deuxième partie de nos contributions, relative à l'intégration de la dimension temporelle dans le processus d'agrégation et d'ordonnancements des documents.
  Ce chapitre présente une nouvelle approche d'agrégation sensible au temps permettant d'adapter les résultats de recherche en fonction des caractéristiques temporelles de la requête. Nous y présentons également une évaluation expérimentale dans le cadre d'une tâche de RI temporelle, en l'occurrence la tâche "Temporel Summarization" de TREC 2013 et 2014.

La troisième partie, intitulée *Conclusion générale* (Chapitre 7) discute l'impact de nos contributions. Ce chapitre conclut cette thèse et présente nos perspectives de recherche.

Première partie

Synthèse des travaux de l'état de l'art

Chapitre 2

Concepts de base de la RI classique

2.1 Introduction

La recherche d'information (RI) est un domaine de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information. Ce besoin en information est souvent formulé en langage naturel par une requête décrite par un ensemble de mots clés. Pour une requête utilisateur, un système de RI permet de retrouver un sous-ensemble de documents susceptibles d'être pertinents, à partir d'une collection de documents, en réponse à cette requête. L'essor du web a remis la RI face à de nouveaux défis d'accès à l'information, à savoir retrouver une information pertinente en tenant compte du cadre de recherche dans lequel se situe l'utilisateur. La problématique majeure de la plupart des moteurs de recherche classiques réside en partie dans le fait qu'ils sont basés sur une approche généraliste qui considère que le besoin en information est complètement représenté par sa requête et délivrant alors des résultats ne tenant compte que de l'adéquation thématique entre les documents et les requêtes. Pour pallier à ces lacunes, des réflexions ont été menées dans le but de mieux cerner la notion de pertinence du point de vue de l'utilisateur et d'identifier les différents facteurs ayant un impact sur cette notion (Borlund, 2003). Les études menées dans ce sens ont montré que la

pertinence n'est pas une relation isolée entre un document et une requête ; elle est définie selon différents niveaux intégrant différents facteurs liés à l'utilisateur et à son environnement dans une situation de recherche d'information (crédibilité et autorité des auteurs, diversité, accessibilité et fraîcheur des résultats de recherche, etc.). Ces dernières proposent des techniques de combinaison de pertinence issues de l'exploitation de plusieurs dimensions de pertinence pour définir un seul score de pertinence des documents.

Ce chapitre traite des concepts de base de la RI classique ainsi que de l'émergence de la notion de pertinence multidimensionnelle. La section 2.2 présente tout d'abord les fondements de la RI classique. Nous abordons les notions et les modèles de base de la RI classique, puis nous présentons la démarche d'évaluation des systèmes de RI. Dans la section 2.3, nous définissons la notion de pertinence et nous détaillons ses différents propriétés. Ensuite, nous montrons l'orientation des travaux vers la RI multicritère pour laquelle nous abordons les problématiques majeures et les verrous scientifiques. La dernière section conclut le chapitre.

2.2 Les fondements de la recherche d'information

La recherche d'information (Rijsbergen, 1979; Salton et McGill, 1986; Baeza-Yates et Ribeiro-Neto, 1999) est la branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la distribution de l'information. En résumé, un système de RI permet de sélectionner à partir d'une collection de documents, des informations pertinentes répondant à des besoins utilisateurs, exprimés sous forme de requêtes. Nous abordons dans la suite de cette section les concepts de base de la recherche d'information, puis nous décrivons le processus général d'un système de RI, ensuite nous passons en revue les principaux modèles de RI et nous présentons la démarche classique d'évaluation des systèmes de RI classique.

2.2.1 Concepts de base de la recherche d'information

Un système de RI est un système qui permet de retrouver, à partir d'une collection de documents, les documents susceptibles d'être pertinents à un besoin en information d'un utilisateur exprimé sous forme d'une requête. Plusieurs concepts clés s'articulent autour de la définition d'un système de RI :

1. *Collection de documents* : la collection de documents (ou *corpus*) constitue l'ensemble des informations (des documents) exploitables et accessibles. Nous utiliserons dans la suite les termes : *corpus* ou collection de manière indifférente.
2. *Document* : le document constitue l'information élémentaire d'une collection de documents. L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document. Nous utiliserons dans la suite les termes *information* ou *document* pour désigner un granule documentaire.
3. *Besoin en information* : cette notion est souvent assimilée au besoin de l'utilisateur. Cosijn et Ingwersen (2000) ont défini trois types de besoins utilisateur :
   - *Besoin vérificatif* : l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un besoin de type vérificatif est dit stable, c'est-à-dire qu'il ne change pas au cours de la recherche.
   - *Besoin thématique connu* : l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et domaine connus. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche.
   - *Besoin thématique inconnu* : pour ce type de besoins, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations hors des sujets ou domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète.
4. *Requête* : la requête est l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le système de RI et l'utilisateur. Divers types de langages d'interrogation sont proposés dans la littérature. Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique.
5. *Pertinence* : la pertinence est une notion fondamentale dans le domaine de la RI. La pertinence peut être définie comme la correspondance entre un document et une requête, ou encore une mesure d'informativité du document à la requête (Boughanem et Savoy, 2008). Borlund (2003) a donné une définition plus large en montrant qu'elle dépend de nombreux critères liés au contexte de la recherche, tels que : le degré de correspondance (*aboutness*), l'utilité (*usefulness/utility*),

rentabilité (*usability*) ou l'importance des résultats retournés par rapport aux objectifs, aux intérêts, à la situation intrinsèque du moment. Ces différents critères ont amené à la catégorisation de la pertinence utilisateur principalement en 5 classes de pertinence : la pertinence algorithmique, la pertinence thématique, la pertinence cognitive, la pertinence situationnelle et la pertinence motivationnelle (ou affective) (Saracevic, 1996). Elles peuvent être définies comme suit :

- *la pertinence algorithmique (ou système)* : souvent présentée par un score de l'adéquation du contenu des documents vis-à-vis de celui de la requête. Pour mesurer cette adéquation, le système de RI procède au calcul du degré de similitude du document et de la requête en se basant sur les représentations internes de chacun de ceux-ci. Le but de tout système de RI est de rapprocher la pertinence algorithmique calculée par le système aux jugements de pertinence donnés par des vrais utilisateurs.
- *la pertinence thématique* : traduit le degré d'adéquation de l'information retrouvée au thème évoqué par le sujet de la requête. C'est la mesure la plus utilisée dans les moteurs de recherche classiques.
- *la pertinence cognitive* : représente la relation entre l'état de la connaissance intrinsèque de l'utilisateur et l'information portée par les documents telle qu'interprétée par l'utilisateur ; cette pertinence se caractérise par une dynamique qui permet d'améliorer la connaissance de l'utilisateur via l'information renvoyée le long de sa recherche.
- *la pertinence situationnelle* : est vue comme l'utilité de l'information retrouvée par rapport à la tâche ou le problème posé par l'utilisateur.
- *la pertinence motivationnelle (ou affective)* : décrit la relation entre les intentions, les buts et les motivations de la recherche tels que fixés par l'utilisateur d'une part et les informations retrouvées d'autre part.

2.2.2 Processus général de la RI

Le but fondamental d'un système de RI est de sélectionner l'ensemble de documents pertinents répondant au besoin en information de l'utilisateur.

La réalisation d'un tel système de RI qui permet, à partir d'une requête, d'ordonnancer les documents consiste principalement à mettre en oeuvre un processus clé (processus en U de la RI). Il est décomposé en trois principales

étapes, illustrées dans la Figure 2.1 et détaillées ci-dessous.

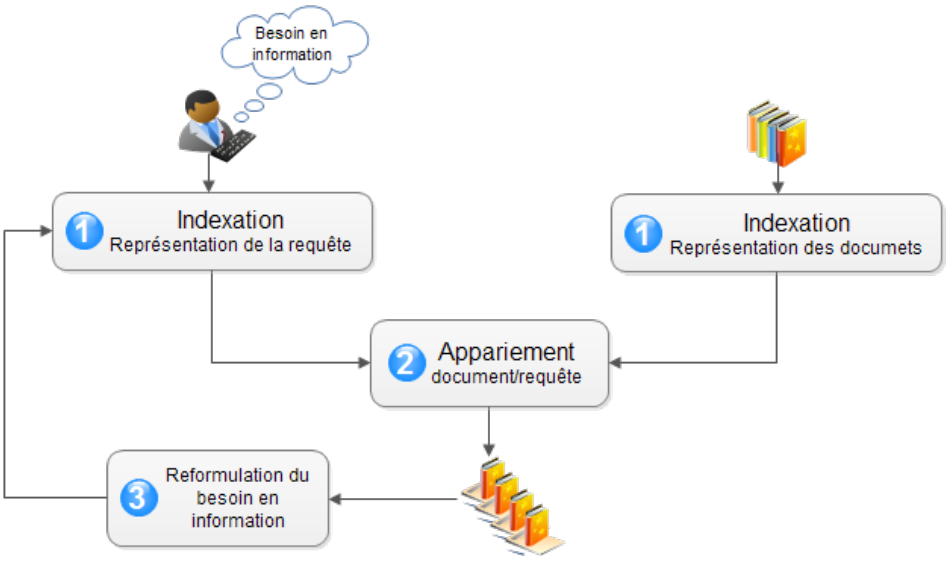


FIGURE 2.1: Processus général de la RI.

Ce processus consiste en deux principales phases : l'indexation et l'interrogation.

1. L'indexation consiste à extraire et à représenter le contenu des documents de manière interne sous forme d'index. Cette structure d'index permet de retrouver rapidement les documents contenant les mots clés de la requête.

2. L'interrogation est l'interaction d'un utilisateur final avec le système de RI, une fois les documents sont représentés sous forme interne d'index. Suite à une requête utilisateur, le système calcule la pertinence de chaque document vis-à-vis de la requête utilisateur selon une mesure de correspondance du modèle de RI, et retourne la liste des résultats à l'utilisateur.

3. La reformulation du besoin en information est l'étape qui permet de redéfinir le besoin de l'utilisateur au fur et à mesure de la session de recherche.

2.2.2.1 La phase d'indexation

L'indexation recouvre un ensemble de techniques visant à transformer les documents (ou requêtes) en substituts ou descripteurs capables de représenter leur contenu (Salton et McGill, 1986). Ces descripteurs forment le langage d'indexation représenté selon une structure souvent basée sur un ensemble de mots clés ou groupes de mots représentant le contenu textuel du document. Dès lors, l'indexation consiste à détecter les termes les plus représentatifs du contenu du document. Différents modes d'indexation existent en RI : l'indexation manuelle, automatique ou semi-automatique.

- *Indexation manuelle* : lors de l'indexation manuelle, un expert dans le domaine choisit les termes qu'il juge pertinents dans la description du contenu sémantique du document. Ce type d'indexation permet d'avoir un vocabulaire d'index contrôlé ce qui permet d'accroître la consistance et la qualité de la représentation obtenue.
- *Indexation automatique* : Ce type d'indexation ne fait pas intervenir d'expert. L'indexation automatique repose sur des algorithmes associant automatiquement des descripteurs à des parties de document. Dans le cas des documents textuels, chaque mot est potentiellement un index du document qui le contient. Chaque terme selon un processus défini : extraction, suppression des mots vides, normalisation et pondération (Porter, 1997; Pirkola et Järvelin, 2001).
- *Indexation semi-automatique* : c'est une combinaison des deux méthodes précédentes où le choix final des termes à indexer revient à l'expert.

A la fin de cette étape, les documents sont représentés dans des fichiers index qui stockent la cartographie des couples terme-document en y associant un poids. La formule de pondération la plus utilisée est celle basée sur la fréquence des termes dans les documents, appelée TF-IDF (Salton et McGill, 1986). L'intuition de cette pondération est de favoriser les termes qui sont à la fois fréquents dans le document et peu fréquents dans la collection. Cette dernière condition est basée sur les propriétés de la loi de Zipf (Zipf, 1949) qui étudie la distribution des termes dans une collection de documents. La mesure TF-IDF est donnée par la multiplication des deux mesures TF et IDF comme suit :

$$TF * IDF = log(1 + TF) * IDF \tag{2.1}$$

Les mesures TF et IDF sont définies comme suit :

1. *TF (Term Frequency)* : cette mesure a été introduite pour tenir compte de la fréquence d'un terme dans un document. L'idée sous-jacente est que plus un terme est fréquent dans un document plus il est important dans sa description. Elle représente une "pondération locale" d'un terme dans un document. On trouve plusieurs variantes de cette mesure. Soit le document d_j et le terme t_i , alors la fréquence TF_{ij} du terme dans le document est donnée selon l'une des formulations suivantes :
$$TF_{ij} = 1 + log(td_{ij}), TF_{ij} = \frac{td_{ij}}{\sum_k td_{kj}} \tag{2.2}$$
où td_{ij} est le nombre d'occurrences du terme t_i dans le document d_j. Le dénominateur est le nombre d'occurrences de tous les termes dans le document d_j. La dernière déclinaison permet de normaliser la fréquence du terme pour éviter les biais liés à la longueur du document.
2. *IDF (Inverse Document Frequency)* : ce facteur mesure la fréquence d'un terme dans toute la collection, c'est la "pondération globale". En effet, un terme fréquent dans la collection, a moins d'importance qu'un terme moins fréquent. Cette mesure est exprimée selon l'une des déclinaisons suivantes :
$$IDF_i = log(\frac{N}{n_i}), IDF_i = log(\frac{N - n_i}{n_i}) \tag{2.3}$$
avec N est la taille (nombre de documents) de la collection et n_i le nombre de documents contenant le terme t_i.

2.2.2.2 La phase d'appariement document-requête

L'interrogation du système implique un processus d'interaction de l'utilisateur avec le système de RI illustré dans la figure 2.1. Cette interaction comprend : (1) la formulation d'une requête par l'utilisateur traduisant son besoin en information ; (2) la représentation de la requête sous forme interne selon le langage d'indexation défini ; et (3) la correspondance entre la requête et les documents par exploitation de l'index et la présentation des résultats. Plus précisément, l'interrogation implique le scénario suivant : l'utilisateur exprime son besoin en information sous la forme d'une requête. Le système interprète la requête et crée son index qui sera compatible avec le modèle d'index des documents. Ensuite, le système évalue la pertinence des documents par rapport à cette requête en utilisant une fonction de correspondance. Cette fonction exploite l'index généré dans la phase d'indexation

dans le but de calculer un score de similarité (en anglais Relevance Status Value), notée RSV (Q, D), entre la requête indexée Q et les descripteurs du document D. Différents modèles de RI ont été proposés dans la littérature tentent de formaliser la pertinence en partant des modèles naïfs basés sur l'appariement exact vers des modèles plus élaborés basés sur l'appariement rapproché. Le résultat est une liste de documents généralement triée par ordre de valeur de correspondance décroissante, c'est-à-dire du plus pertinent au moins pertinent, et présenté à l'utilisateur. Celui ci apporte son jugement sur les documents renvoyés par le système selon des critères liés à son besoin en information et au cadre de recherche dans lequel il se situe.

2.2.2.3 La phase de reformulation du besoin en information

La reformulation du besoin en information est l'étape qui permet de redéfinir le besoin de l'utilisateur au fur et à mesure de la session de recherche. Cette étape peut être effectuée :
- Manuellement, dans le cas où l'utilisateur soumet lui-même une nouvelle requête.
- De façon automatique, lorsque le système de RI s'appuie sur les termes importants dans les documents les plus pertinents ou visités par l'utilisateur qui sont réutilisés.

2.2.3 Aperçu des principaux modèles de RI

Un modèle de RI se définit principalement, par sa modélisation de la mesure de la pertinence document-requête, mais aussi par sa représentation des documents et sa représentation des requêtes. Une taxonomie des modèles a été présentée par (Baeza-Yates et Ribeiro-Neto, 1999) et présente quatre familles principales.

Comme illustrée dans la Figure 2.2, les modèles reposent sur le texte des documents (modèles de RI classiques et modèles basés sur le texte semi-structuré), les liens entre les documents (modèles orientés web) et les documents multimédia (recherche d'images, de musiques, d'audio ou de vidéos). Compte tenu des concepts utilisés dans nos contributions, nous présentons dans cette sous-section les modèles appartenant à la catégorie des modèles de RI classiques reposant respectivement sur la théorie des ensembles, les méthodes algébriques et les probabilités. Pour ce faire, nous considérons les notations suivantes :

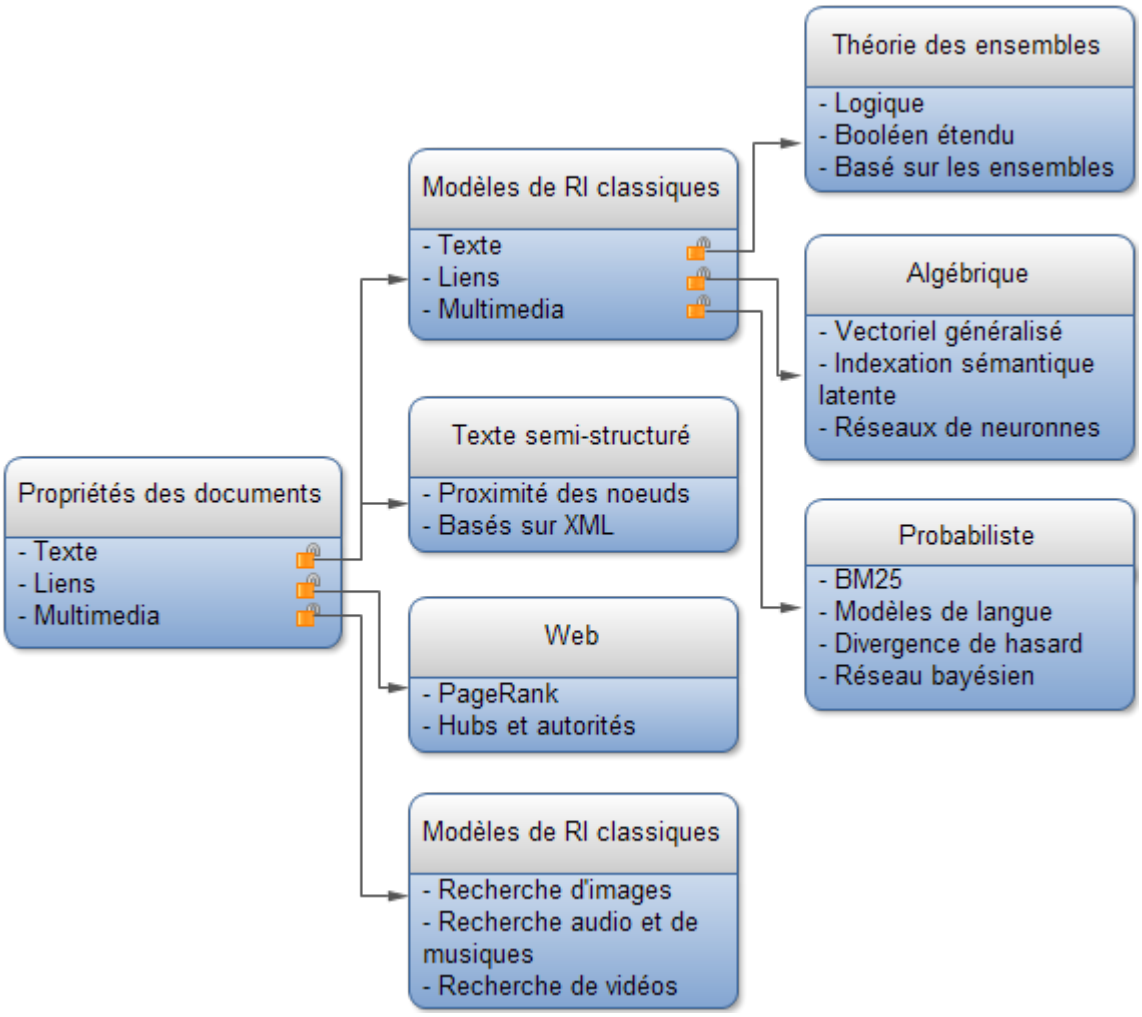


FIGURE 2.2: Taxonomie des modèles de RI (Baeza-Yates et Ribeiro-Neto, 2011)

- L'index I est modélisé par le vecteur $I = (t_1, \ldots, t_v, \ldots, t_V)$ où chaque élément t_v représente un terme de l'index I et V correspond au nombre de termes dans l'index.
- La collection de documents est notée $D = \{d_1, \ldots, d_i, \ldots, d_N\}$ où N représente le nombre de documents dans la collection.
- L'ensemble de requêtes est noté $Q = \{q_1, \ldots, q_h, \ldots, q_Z\}$ où Z représente le nombre de requêtes.
- Le document $d_i \in D$ est modélisé vecteur de poids $d_i = (w_{i_1}, \ldots, w_{i_v}, \ldots, w_{i_V})$ où chaque élément w_{i_v} représente le poids du terme t_v pour le document d_i.
- La requête $q_h \in Q$ est modélisée par un vecteur $q_h = (w_{h_1}, \ldots, w_{h_v}, \ldots, w_{h_V})$ où chaque élément w_{h_v} représente le poids du terme t_v pour la requête q_h.
- La fonction d'appariement est notée $RSV(d_i, q_h)$ et retourne le score de similarité du document d_i par rapport à la requête q_h.

2.2.3.1 Modèle booléen

Le modèle booléen (Salton, 1968) est basé sur la théorie des ensembles. Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés. Chaque document est représenté par une conjonction logique des termes non pondérés qui constitue l'index du document. Un exemple de représentation d'un document est comme suit : $d = t_1 \wedge t_2 \wedge t_3 ... \wedge t_n$.

Une requête est une expression booléenne dont les termes sont reliés par des opérateurs logiques (OR, AND, NOT) permettant d'effectuer des opérations d'union, d'intersection et de différence entre les ensembles de résultats associés à chaque terme. Un exemple de représentation d'une requête est comme suit : $q = (t_1 \wedge t_2) \vee (t_3 \wedge t_4)$. La fonction de correspondance est basée sur l'hypothèse de présence/absence des termes de la requête dans le document et vérifie si l'index de chaque document d implique l'expression logique de la requête q. Le résultat de cette fonction, décrite comme : $RSV(q, d) = 1, 0$, est binaire.

2.2.3.2 Modèle vectoriel

Dans ces modèles, la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. Le modèle vectoriel (Salton, 1971) représente les documents et les requêtes par des

vecteurs d'un espace à n dimensions, les dimensions étant constituées par les termes du vocabulaire d'indexation. L'index d'un document d_j est le vecteur $\vec{d_j} = (w_{1,j}, w_{2,j}, w_{3,j}, \ldots, w_{n,j})$, où $w_{k,j} \in [0,1]$ dénote le poids du terme t_k dans le document d_j. Une requête est également représentée par un vecteur $\vec{q} = (w_{1,q}, w_{2,q}, w_{3,q}, .., w_{n,q})$, où $w_{k,q}$ est le poids du terme t_k dans la requête q. La fonction de correspondance mesure la similarité (l'angle) entre le vecteur requête et les vecteurs documents. Il existe à ce jour plusieurs mesures dont les plus connues sont les suivantes :

– *Le produit scalaire* :

$$RSV(\vec{q}, \vec{d_j}) = cos(\vec{q}, \vec{d_j}) \tag{2.4}$$

– *La mesure de cosinus* où $RSV(q, d_j) = \frac{\vec{q} \cdot \vec{d_j}}{||\vec{q}|| \cdot ||\vec{d_j}||}$ où $||\vec{x}||$ représente la norme euclidienne du vecteur $\vec{x}$.
– *La mesure de Jaccard* où $RSV(q, d_j) = \frac{|\vec{q} \cap \vec{d_j}|}{|\vec{q} \cup \vec{d_j}|}$ où $|\vec{q} \cap \vec{d_j}|$ correspond au nombre de termes présents à la fois dans la requête q et le document d_j tandis que $|\vec{q} \cup \vec{d_j}|$ représente le nombre de termes contenus dans la requête q ou le document d_j
– *La mesure de Dice* où $RSV(q, d_j) = \frac{2 \cdot |\vec{q} \cap \vec{d_j}|}{|\vec{q}| + |\vec{d_j}|}$ où $|\vec{d_j}|$, respectivement $|\vec{q}|$, annote le nombre de termes dans le document d_j, respectivement la requête q.

A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne satisfont la requête qu'approximativement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante. L'inconvénient de ce modèle est qu'il repose sur l'hypothèse d'indépendance des termes –*bag of words*– alors que ce sont parfois les expressions ou les groupes de mots qui enrichissent la sémantique du document. Une des réponses à ce problème réside dans la considération des N-grammes (Song et Croft, 1999), permettant de regrouper des termes successifs qui peuvent avoir du sens ensemble.

2.2.3.3 Modèle probabiliste

Ce modèle est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête (Maron et Kuhns, 1960; Robertson et Jones, 1976; Salton et McGill, 1986). Nous distinguons deux principales catégories de

modèles probabilistes : les modèles probabilistes classiques et les modèles de langues.

Les modèles probabilistes classiques. Ces modèles (Robertson et Walker, 1994; Robertson *et al.*, 1995) reposent sur la distribution de probabilité des termes pour identifier la similarité document-requête. Le principe de ce modèle (Robertson *et al.*, 1995) est de favoriser les documents à la fois caractérisés par une forte probabilité d'être pertinents (événement P) et une faible probabilité d'être non pertinent (événement $\overline{P}$). Le score de pertinence d'un document d_j par rapport à la requête q est estimé comme suit :

$$RSV(q, d_j) = \frac{P(P|d_i)}{P(\overline{P}|d_j)} \tag{2.5}$$

où $P(P|d_j)$, respectivement $P(\overline{P}|d_j)$, représente la probabilité de pertinence, de non pertinence, par rapport à la requête q compte tenu du document d_j. Cette fonction d'appariement peut être estimée ainsi :

$$RSV(q, d_j) = \prod_{t_v \in q} \frac{p_v(1 - qv)}{q_v(1 - p_v)} \tag{2.6}$$

avec $p_v = \frac{r_v}{n}$ et $q_v = \frac{R_v - r_u}{N - n}$, et où $t_v \in q$ correspond à l'ensemble des termes t_v de la requête q. p_v et q_v représentent respectivement la probabilité que le terme t_v apparaisse dans le document d_j sachant qu'il est pertinent, respectivement non pertinent, par rapport à la requête. Ces probabilités sont estimées par maximum de vraisemblance sur l'ensemble de la collection D et dépendent du nombre total R de documents pertinents, dont r_v documents contenant t_v, ainsi que le nombre total N de documents dans la collection incluant n documents pertinents. Après développement et en ajoutant un coefficient de 0.5, afin d'éviter de diviser par 0, la formule finale, est la suivante :

$$RSV(q, d_j) = \prod_{t_v \in q} \frac{(r_v + 0.5)(N - n_v - R + r_v + 0.5)}{(n_v - r_v + 0.5)(R - r_v + 0.5)} \tag{2.7}$$

De nombreuses applications du modèle probabiliste ont été proposées dans la littérature, telles que le Okapi BM25 (Robertson *et al.*, 1995) ou le modèle binaire BIR (Yu et Salton, 1976). Le modèle le plus utilisé est le modèle Okapi BM25. Les atouts majeurs de ce modèle consistent en la considération de la longueur des documents dans le calcul de la pertinence et de la fréquence des termes dans la collection, conformément à la loi de (Zipf, 1949).

La fonction d'appariement est présentée dans l'équation ci-dessous :

$$RSV(q, d_j) = \sum_{t_v \in q} \frac{(N - n_v) + 0.5}{n_k + 0.5} \frac{f_{iv} \cdot (k_1 + 1)}{f_{iv} + k_1 \cdot (1 - b + b \cdots \frac{|d_i|}{avg_{dl}})} \tag{2.8}$$

où N représente la taille de la collection, n_v le nombre de documents qui contiennent le terme t_v. La fréquence du terme t_v dans le document d_i est notée f_{iv}. $|d_j|$ représente la longueur du document d_j tandis que la longueur moyenne des documents est notée avg_{dl}. Deux paramètres, respectivement k_1 et b sont utilisés et ont obtenu par expérimentation les valeurs optimales suivantes : $k_1 \in [1.2; 2.0]$ et $b = 0.75$.

Les modèles de langue. Le principe des modèles de langue (Ponte et Croft, 1998) repose sur le fait que la pertinence d'un document estime la similarité entre la requête et le modèle de langue du document θ_d. Le score de similarité $RSV(q, d_j)$ est calculé comme suit :

$$RSV(q, d_j) = P(q|\theta_d) = \prod_{t_v \in q} P(t_v|\theta_{d_j}) \tag{2.9}$$

où $P(q|\theta_{d_i})$ représente la probabilité de la requête q sachant le modèle de langue θ_{d_i} du document d. Pour chaque terme t_v appartenant à la requête q, sa probabilité par rapport au modèle de langue θ_{d_j} du document d est notée $P(q|\theta_{d_j})$. Cette dernière probabilité s'appuie sur une estimation de la fréquence des termes de la requête q dans le document d mais est annulée pour les documents ne contenant pas tous les termes de la requête. Dans ce cas particulier, le score de similarité du document est nul alors que le document pourrait partiellement répondre au besoin en information formulé par la requête. Pour pallier cet inconvénient, des techniques de lissage ont été proposées (Jelinek et Mercer, 1980; MacKay et Peto, 1994; Chen et Goodman, 1996). Ces dernières s'appuient sur un modèle de référence, en l'occurrence celui de la collection, pour estimer la pertinence d'un terme sur ce modèle de référence.

2.2.4 Évaluation des performances des systèmes de RI

L'évaluation d'un système de RI permet de vérifier l'efficacité des modèles mis en œuvre pour l'identification des documents pertinents. Dans cette

section, nous présentons le cadre d'évaluation d'un système de RI ainsi que les mesures d'évaluation sous-jacentes.

2.2.4.1 Collections de test

La collection de test constitue le contexte d'évaluation, c'est-à-dire les éléments qui vont servir à évaluer un modèle de RI. Une collection de test regroupe un ensemble de documents, des requêtes formulant des besoins en information et des jugements de pertinence associés qui recensent les documents pertinents pour chacune des requêtes. Cette approche d'évaluation correspond au paradigme de Cranfield (Cleverdon, 1997) qui a suscité le développement de nombreuses campagnes d'évaluation depuis les années 1970. Ces dernières présentent l'avantage de cibler une tâche particulière et d'évaluer l'efficacité des systèmes répondant à cette tâche. Les campagnes d'évaluation les plus connues sont :

1. La campagne TREC [1] –Text REtrieval Conference– est une des premières des campagnes qui regroupe à ce jour un large panel de tâches, telles que la recherche *ad-hoc*, ou également les tâches de recherche dans les microblogs ou celles orientées pour les systèmes de questions-réponses.
2. La campagne INEX –Initiative for the Evaluation of XML Retrieval– oriente ses tâches de recherche vers des collections de documents structurés.
3. La campagne CLEF [2] –Conference and Labs of the Evaluation Forum– propose des campagnes dans des langues différentes de l'anglais, traité majoritairement dans les campagnes TREC. En plus de proposer des tâches de recherche sur des documents, cette campagne fournit également des collections d'images associées à des annotations.
4. La campagne NTCIR [3] a aussi développé diverses collections d'essais, avec une attention particulière aux langues d'Asie de l'Est et la recherche d'information multilingue. Les requêtes sont faites dans une langue, toutefois, les collections de documents contiennent des documents dans une ou plusieurs autres langues. Cette campagne propose différentes tâches d'évaluation de système de RI telles que les systèmes

1. `http://trec.nist.gov`
2. `http://www.clef-initiative.eu/`
3. `http://research.nii.ac.jp/ntcir/ntcir-12/tasks.html`

de questions-réponses (Q&A task), RI mobile (tâche MobileClick-2), RI temporelle, etc.

Dans TREC, les recherches étaient centrées au départ (de TREC 1 à TREC 6) sur deux tâches principales : la tâche de routage et la tâche *ad-hoc*. La tâche *ad-hoc* est constituée d'un ensemble de nouvelles requêtes qui sont lancées sur une collection de documents fixés, et la tâche de routage est composée d'un ensemble de requêtes fixes lancées sur une collection de documents en évolution. L'émergence de la RI orienté utilisateur a recentré ce dernier au sein du processus d'évaluation. De nouvelles tâches considérant la dimension de l'utilisateur sont apparues, parmi lesquelles :

1. *La tâche TREC Interactive* : qui consiste en la résolution d'un besoin complexe. Les participants doivent alors fournir les fichiers log qui recensent les interactions des utilisateurs (requêtes soumises, documents visités, ...).
2. *La tâche TREC Contextual Suggestion* : qui consiste en une suggestion de documents à partir d'un ensemble de profils utilisateur et d'un contexte, traduisant respectivement les préférences et la localisation des utilisateurs.
3. *La tâche TREC Session Search* : qui consiste en l'ordonnancement des documents vis-à-vis d'une requête particulière, soumise à un moment donné de la session, à partir de l'historique de recherche antérieur d'un utilisateur (requêtes reformulées antérieurement et leurs ordonnancements et jugements de pertinence associés).

Chacune de ces tâches d'évaluation propose une ou plusieurs collections de tests, généralement composées : d'une collection de documents, d'une collection de requêtes, et des jugements de pertinence des documents par rapport à ces requêtes.

1. *Collection de requêtes* : appelées aussi "*topics*", simule l'activité de recherche de l'utilisateur. Pour exploiter au mieux les caractéristiques de la collection de documents et avoir une évaluation assez objective, il est important de créer un ensemble de quelques dizaines de requêtes et qui soient adéquates par leur longueur, les thèmes abordés, leur forme, etc. Les requêtes sont généralement artificielles formulées par des assesseurs qui participent à la compagne d'évaluation, mais elle peuvent aussi être de vraies requêtes extraites à partir de log de recherche Web comme c'est le cas pour la tâche Web de TREC.

2. *Corpus de documents* : c'est un ensemble de documents sur lesquels les systèmes de RI posent des requêtes et récupèrent les documents pertinents. Il existe de très nombreux ensembles de documents en accès libre, notamment sur le Web : des documents plus ou moins vulgarisés, plus ou moins spécialisés dans un domaine, dans une langue ou une autre, etc. Le choix d'une collection ou autre dépend de la tâche de recherche que l'on veut évaluer, pour garantir une représentativité par rapport à la tâche. De même que la spécification du volume des collections de documents utilisées dans l'évaluation est relativement dépendante de la tâche de recherche impliquée dans le système de RI à évaluer, pour garantir une diversité des sujets et du vocabulaire.
3. *Jugements de pertinence* : Les jugements de pertinence indiquent pour chaque document du corpus s'il est pertinent, et parfois même à quel degré il l'est, pour chaque requête. Pour établir ces listes de documents pour toutes les requêtes, les utilisateurs (ou des testeurs simulant des utilisateurs) doivent examiner chaque document de la base de documents, et juger s'il est pertinent par rapport à une requête donnée. Dans les programmes d'évaluation tels que TREC, les collections de documents contiennent plus d'un million de documents, ce qui rend impossible le jugement exhaustif de pertinence. Ainsi, dans le cas de grandes collections, les jugements de pertinence sont construits selon la technique de *pooling*, effectuée à partir des 1000 premiers documents retrouvés par les systèmes participants. Cette technique est souvent utilisée dans les compagnes d'évaluation telles que TREC ou CLEF.

2.2.4.2 Protocole d'évaluation

Le protocole d'évaluation dans le modèle d'évaluation orienté-laboratoire définit une méthodologie rigoureuse et efficace pour comparer plusieurs systèmes de RI, stratégies de recherche, ou algorithmes sur une même base, en spécifiant trois composants non indépendants qui sont : le nombre de topics utilisés, les mesures d'évaluation utilisées et la différence de performances requises pour considérer qu'une stratégie de recherche est meilleure qu'une autre (Buckley et Voorhees, 2000). L'évaluation de l'efficacité de chaque stratégie de recherche consiste à évaluer la liste des résultats obtenus pour chaque requête de test. Cette évaluation est à la base de la correspondance entre la pertinence algorithmique calculée par le système et la pertinence donnée par les assesseurs. L'efficacité globale d'une stratégie de recherche est calculée comme étant la moyenne des précisions calculées selon une me-

sure donnée sur l'ensemble des topics dans la collection de test. Les protocoles d'évaluation utilisés se basent sur des métriques, nous présentons dans la suite les mesures les plus courantes dont les plus classiques "Rappel" et "Précision".

2.2.4.3 Mesure d'évaluation

Les mesures d'évaluation permettent d'estimer quantitativement l'efficacité d'un système. L'objectif est d'identifier, pour chaque requête la capacité du système à retourner des documents pertinents. La principale difficulté d'un système de RI est de reposer sur un modèle qui retourne le maximum de documents pertinents sans augmenter le nombre de documents non pertinents retournés. Chaque requête q est évaluée individuellement au moyen d'une mesure statistique estimée au rang r de la liste l retournée par le système de RI. La mesure est ensuite agrégée sur l'ensemble des requêtes de la collection de test afin d'obtenir la mesure d'efficacité moyenne du système.

Nous présentons dans la suite les mesures les plus classiques de "Rappel" et de "Précision", ainsi qu'un ensemble de mesures les plus courantes.

1. *Rappel et précision* : étant donnée une requête q, les documents de la collection peuvent être classifiés en fonction de leur rapport à la requête (pertinents/non pertinents). On considère $|S|$ le nombre de documents sélectionnés par un système de RI pour la requête q. On considère de plus, le nombre $|P|$ des documents pertinents dans la collection pour cette requête et $|PS|$ le nombre des documents pertinents sélectionnés par le système. La mesure de précision calcule alors la capacité du système à rejeter tous les documents non pertinents pour une requête. Elle est donnée par le rapport entre les documents sélectionnés pertinents et l'ensemble des documents sélectionnés :

$$Précision = \frac{|PS|}{|S|} \tag{2.10}$$

   Le rappel mesure la capacité du système à renvoyer tous les documents pertinents pour une requête. Il est donné par le rapport entre les documents pertinents sélectionnés et l'ensemble des documents pertinents pour la requête :

$$Rappel = \frac{|PS|}{|P|} \tag{2.11}$$

La précision mesurée indépendamment du rappel (et inversement) est peu significative. En pratique, les valeurs du rappel et de précision sont conjointement calculées à chaque document restitué pour les i premiers documents dans la liste des réponses du système. Ces deux mesures évoluent en sens inverse. Intuitivement, si on augmente le rappel en retrouvant plus de documents pertinents, on diminue la précision en retrouvant aussi plus de documents non pertinents. Inversement, une plus grande précision risque de rejeter des documents pertinents diminuant ainsi le rappel. Le comportement d'un système peut varier en faveur de la précision ou en faveur du rappel au détriment de l'autre métrique.

2. *F-mesure* : c'est une mesure qui combine la précision et le rappel, nommée F-mesure ou F-score introduite dans (Rijsbergen, 1979) et définie par :

$$F_\beta = \frac{(1+\beta^2)\cdot(\text{précision}\cdot\text{rappel})}{(\beta^2\cdot\text{précision}+\text{rappel})} \tag{2.12}$$

pour des valeurs réelles positives de β traduisant l'importance relative du rappel et de la précision. Un cas particulier de la mesure générale F_β est comme par la mesure $F1$ ($\beta = 1$), dans ce cas particulier la F-mesure définit la moyenne harmonique du rappel et de la précision :

$$F = \frac{2\cdot\text{précision}\cdot\text{rappel}}{\text{précision}+\text{rappel}} \tag{2.13}$$

3. *précision@X* : c'est la précision à différents niveaux de coupe. Cette précision mesure la proportion des documents pertinents retrouvés parmi les X premiers documents retournés par le système.
4. *R-précision* : cette précision mesure la proportion des documents pertinents retrouvés après que R documents ont été retrouvés, où R est le nombre de documents pertinents pour la requête considérée.
5. *Précision moyenne (Mean Average Precision)* : c'est la moyenne des précisions moyennes (Average precision-AP) obtenues sur l'ensemble des requêtes à chaque fois qu'un document pertinent est retrouvé :

$$MAP = \frac{\sum_{q\in Q} AP_q}{|Q|} \tag{2.14}$$

avec AP_q est la précision moyenne d'une requête q, Q est l'ensemble des requêtes et $|Q|$ est le nombre de requêtes. Cette mesure peut être

qualifiée de globale puisqu'elle combine différents points de mesure. Cette précision peut être aussi calculée à différents niveaux de rappel $(0\%, 10\%, 20\%, ..., 100\%)$, elle est alors appelée : précision moyenne interpolée (MAiP). Pour chaque niveau de rappel, les valeurs calculées sont moyennées sur tout l'ensemble des requêtes.

6. *La mesure BPREF* : dans le cas de collections volumineuses, la construction de jugements de pertinence complets est diffcile voir impossible puisque elle est très coûteuse en terme de temps. Afin de pallier cet inconvénient, (Buckley et Voorhees, 2000) ont proposé la mesure BPREF (Binary PREFerence-based measure). Cette mesure ne considère que les documents jugés et elle prend en compte les documents pertinents et les documents non pertinents. Elle est donnée par la formule suivante :

$$bpref = \frac{1}{R}\sum_{r} 1 - \frac{n \text{ classés avant } r}{R} \tag{2.15}$$

Avec R le nombre de documents pertinents pour la requête, r est un document pertinent et n est le nombre de documents non pertinents classés avant le document pertinent r.

7. *Mean Reciprocal Rank (MRR)* : une autre mesure basée rang est la métrique Mean Reciprocal Rank. Elle permet d'évaluer le nombre de documents qu'il faut considérer avant de retrouver le premier document pertinent. Elle est égale à la moyenne calculée sur l'ensemble des requêtes, du rang du premier document pertinent.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{Q} \frac{1}{\text{rank}_i} \tag{2.16}$$

MRR est nulle pour une requête si aucun document pertinent n'est retourné par le système. Cependant, MRR donne un score élevé pour un système qui retourne des documents pertinents en haut de la liste présentée à l'utilisateur. Cette mesure est couramment utilisée dans les systèmes Questions-Réponses où l'utilisateur s'intéresse à recevoir la bonne réponse en premier rang.

2.3 De la pertinence thématique à la pertinence multidimensionnelle

La littérature concernant le domaine de la RI a connu un très grand nombre de publications portant sur le concept de pertinence durant les deux dernières décennies (Saracevic *et al.*, 1974; Schamber, 1991; Barry, 1994; Mizzaro, 1998; Cosijn et Ingwersen, 2000; Saracevic, 2007b). La dimension de pertinence la plus couramment utilisée est la pertinence thématique (ou aussi topicale (Vickery, 1959)) qui traduit la proximité des sujets des requêtes et de documents. Plusieurs travaux ont souligné que ce critère thématique est la dimension principale de pertinence et que tous les autres critères en sont dépendants (Saracevic, 2007b). Cependant, un nombre considérable d'études ont souligné l'aspect multidimensionnel de ce concept (Saracevic *et al.*, 1974; Schamber, 1991; Barry, 1994; Cosijn et Ingwersen, 2000; Borlund, 2003; Duan *et al.*, 2010; Saracevic, 2007b). Dans cette section, nous présentons les différents travaux liés à la modélisation de pertinence et nous abordons les problématiques majeures et les verrous scientifique liés à l'estimation de la pertinence multidimensionnelle.

2.3.1 Notion de pertinence multidimensionnelle

Le concept de pertinence est incontestablement au centre d'une activité de recherche d'information comme en témoignent les nombreux travaux qui en ont fait l'objet d'étude (Saracevic, 1976; Borlund, 2003; Saracevic, 2007b). L'un des résultats phares qui ressort de ces travaux est que la pertinence est estimée en globalité selon un ensemble de dimensions qui s'apparentent à des familles de critères ; parmi ces différentes dimensions, on cite les plus reconnues dont : la pertinence thématique (contenu et méta-contenu), la pertinence situationnelle (temps et géolocalisation) et la pertinence cognitive (expertise, centres d'intérêts). Un autre résultat important est l'interdépendance de ces dimensions pour inférer la pertinence globale d'un document (Nagmoti *et al.*, 2010; Saracevic, 2007b). En clair, un utilisateur juge de la pertinence d'un document en tenant compte conjointement de l'ensemble des critères de pertinence ; à titre d'exemple, un document est d'autant plus pertinent du point de vue du contenu que l'expertise de l'utilisateur est en lien avec ce contenu. Le Tableau 2.1 recense l'ensemble des travaux qui ont modélisé et exploité le concept pertinence.

Les travaux respectifs de (Cuadra et Katter, 1967) et (Rees et Schultz, 1967)

Références	**Critères de pertinence étudiés**
(Cuadra et Katter, 1967; Rees et Schultz, 1967)	40 critères comprenant le style et le niveau de difficulté du document.
(Cooper, 1973)	Nouveauté (Novelty), informativeness, crédibilité, importance, clarté, facteurs positifs/négatifs
(Taylor, 1986)	Facilité d'utilisation (Ease of use), réduction du bruit, qualité, adaptabilité, gain de temps
(Schamber, 1991)	10 critères (3 catégories ; *information, source, présentation*)
(Su, 1992, 1994)	20 mesures (groupes : *succès, efficacité, utilité, satisfaction de l'utilisateur*)
(Barry, 1994)	24 Critères groupés en 7 classes
(Saracevic, 1996)	Pertinence (Topicale, algorithmique, cognitive, situationnelle, motivationnelle/affective)
(Mizzaro, 1998)	Ressources d'informations, problèmes utilisateurs, temps, composants
(Cosijn et Ingwersen, 2000)	Pertinence topicale, cognitive, situationnelle, socio-cognitive
(Borlund, 2003)	Pertinence topicale, cognitive, situationnelle

TABLE 2.1: Synthèse des travaux liés à la pertinence multidimensionnelle.

ont analysé les facteurs impactant la perception de pertinence des utilisateurs et ont identifié 40 variables possibles qui peuvent influencer le jugement des documents. Dans la même direction de recherche, Barry (1994) a mené une étude exploratoire dans laquelle elle a repéré 23 catégories de pertinence. Ces catégories incluent plusieurs critères parmi lesquels des dimensions liées au contexte de l'utilisateur (situation, environnement) et d'autres liées à la qualité du document source (autorité et réputation). Plus tard, les recherches sur la pertinence se sont focalisées sur d'autres aspects qui sont plutôt liés à l'utilisateur (Cosijn et Ingwersen, 2000; Borlund, 2003). A titre d'exemple, (Borlund, 2003) a souligné l'importance de 3 dimensions de per-

tinence : *topicale*, *cognitive* et *situationnelle*. De nombreux autres travaux en RI ont aussi mis en exergue à la fois l'importance et la complexité du concept pertinence (da Costa Pereira *et al.*, 2009, 2012; Gerani *et al.*, 2012), mais ces derniers se sont focalisés sur la modélisation d'approches théoriques pour la combinaison des critères de pertinence potentiels identifiés.

2.3.2 Facteurs d'émergence des approches multicritères pour la RI

La limite majeure de la RI classique est qu'elle est basée sur une approche généraliste qui évalue invariablement les requêtes des utilisateurs et délivrent des résultats sans tenir compte des différents critères pouvant impacter le jugement de pertinence des utilisateurs tels que ses préférences ou son environnement de recherche (localisation, dispositif de recherche). La figure 2.3 montre un exemple des différents critères qui peuvent impacter les jugement de pertinence selon le contexte de RI donné.

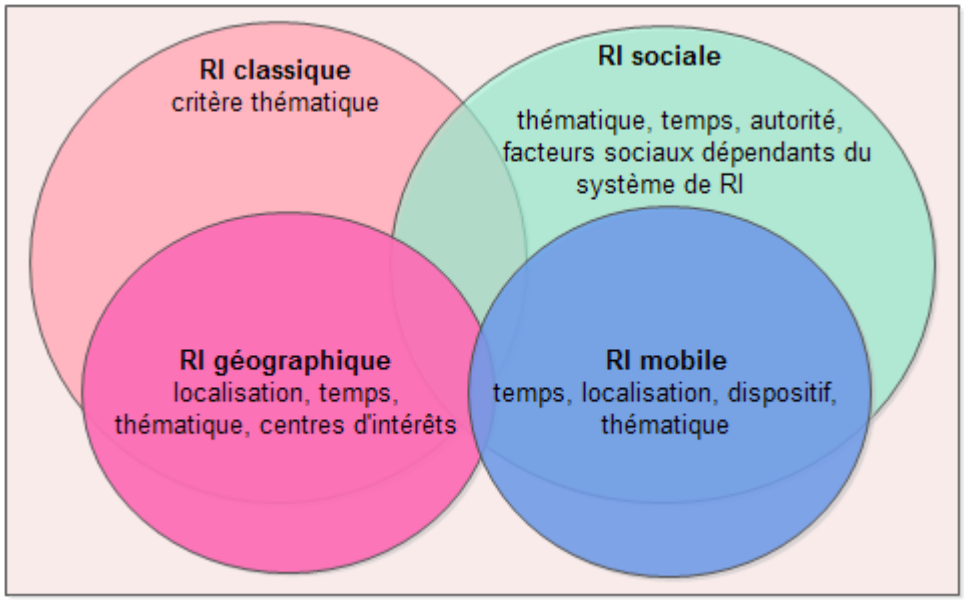


FIGURE 2.3: Un exemple de cadre de RI avec les différents facteurs de pertinence associés.

Par exemple, dans un contexte de RI géographique, un utilisateur en déplacement qui émet la requête "*restaurant pas cher*" s'intéresse plus aux restau-

rant proches de sa localisation que ceux qui sont dans une autre zone géographique. Donc le critère emplacement est très important dans ce contexte de recherche. Un système de RI idéal doit inclure cette dimension de pertinence dans son évaluation des restaurants à retourner. Le critère "prix" (*pas cher*) est aussi important pour l'utilisateur. De plus, si le système de RI dispose des données précédentes (historique de recherche) de l'utilisateur, il pourrait tenir compte de ses préférences dans le classement des résultats (e.g., s'il préfère des restaurants touristiques, café-restaurant, etc). Par ce type de requête, l'objectif principal de la RI peut être formulé comme suit : le système de RI se propose de choisir à partir d'un ensemble de documents ceux répondant au mieux aux différents critères d'un utilisateur. Cette problématique est traitée par les approches de RI multicritères. L'émergence de ces méthodes est principalement dûe à la la prolifération des ressources d'information hétérogènes (blogs, tweets), la diversité des besoins en informations des utilisateurs ainsi que l'apparition de plusieurs cadres de RI (sociale, mobile, géographique) qui permettent aux systèmes de RI d'exprimer le besoin en information en fonction de plusieurs facteurs.

2.3.3 Verrous scientifiques

L'importance de la notion de pertinence est liée au fait que la notion sous-jacente est le fondement des modèles d'ordonnancement de documents en réponse à une requête, qui est la finalité même d'un système de RI (Baeza-Yates et Ribeiro-Neto, 1999). Sa complexité est, quant à elle, subordonnée à plusieurs propriétés :

1. *Multiplicité des dimensions* : qui peuvent être de surcroît, interdépendantes ; même si de nombreux travaux du domaine se sont focalisés sur la dimension thématique seule, force est de constater que de nombreux autres travaux ont prouvé empiriquement l'impact conjoint de plusieurs dimensions sur l'estimation de la pertinence finale, comme la tâche et la situation de recherche (Borlund, 2003; Saracevic, 2007b; Taylor *et al.*, 2007). Considérons à titre d'exemple, une tâche de recherche de tweets ; des analyses expérimentales ont montré que la pertinence d'un tweet en réponse à une requête, est impactée principalement par la conjonction de trois dimensions qui sont le sujet et la fraîcheur du tweet et l'autorité du *tweeterer* qui l'a émis (Nagmoti *et al.*, 2010).
2. *Subjectivité qui entoure les dimensions* : la plupart d'entre elles ne sont pas basées sur des estimations objectives puisqu'elles sont fortement

liées à la perception personnelle des utilisateurs impliqués dans la tâche de RI ; on cite à titre d'exemple les centres d'intérêt, l'expertise et les préférences des utilisateurs.

3. *Modélisation de la pertinence* : définir des opérateurs capables d'agréger des scores de pertinence partiels (relatifs à chaque dimension) en tenant compte de leur interdépendance éventuelle. Cette problématique a été abordée dans diverses applications de RI comme la RI personnalisée (Sieg *et al.*, 2007; Daoud *et al.*, 2010), la RI mobile (Göker et Myrhaug, 2008), la RI sociale (Nagmoti *et al.*, 2010) et la RI géographique (Mata et Claramunt, 2011). Cependant, ces travaux applicatifs ont généralement utilisé des opérateurs de calcul de moyenne pondérée ou de combinaison linéaire qui se basent sur l'hypothèse non réaliste d'additivité ou d'indépendance des dimensions. D'autres travaux fondamentaux récents, se sont intéressés en revanche à la définition d'opérateurs d'agrégation, indépendamment du cadre applicatif, qui permettent de traiter peu ou prou le biais de l'interactivité (da Costa Pereira *et al.*, 2012; Gerani *et al.*, 2012; Eickhoff *et al.*, 2013a). Toutefois, ces opérateurs ne permettent pas de tenir compte de la propriété de subjectivité qui peut se décliner à travers les différences entre les utilisateurs quant à l'importance accordée à chaque dimension de pertinence.

4. *Évaluation* : à ce jour, l'agrégation de pertinence multicritère ne dispose pas de cadre formel pour l'expérimentation. Aucune collection de test officielle dédiée, composée d'un ensemble de documents, de requêtes tests et de jugements de pertinence associés à tous les critères, n'a été constituée pour l'évaluation de cette tâche particulière. Bien qu'il existe des collections existantes à partir desquelles on peut identifier plusieurs critères, ces dernières on été évalués à la base de la pertinence globale des documents. Par exemple, nous pouvons identifier plusieurs critères de pertinence à partir de la collection de test standard de la tâche Microblog de TREC, mais nous disposons pas d'évaluations associées aux facteurs éventuellement identifiés. Les assesseurs de NIST qui jugent la pertinence des documents retournés par les participants à la tâche, donnent des jugements uniquement suivant la pertinence globale des tweets. Afin de valider les modèles de combinaison de pertinence multicritère, la formalisation d'un cadre expérimental et la constitution de collections de test formelles deviennent un enjeu important. En outre, la plupart des mesures de test utilisées sont des mesures adaptées uniquement à des tâches de recherche classiques

n'impliquant que la pertinence thématique.

2.4 Conclusion

Nous avons présenté dans ce chapitre les concepts de base et les principaux modèles de la RI classique. Nous avons également donné un aperçu des différents travaux effectués sur la notion de pertinence puis nous avons cité quelques limitations des approches de RI classiques pour l'estimation de ce concept. Nous avons aussi donné les facteurs d'émergence des approches multicritères pour la RI et nous avons montré l'orientation des travaux vers la RI multicritère. Dans le chapitre suivant, nous apportons un aperçu des approches multicritères pour l'estimation de la pertinence multidimensionnelle et nous présentons les différentes méthodes d'agrégation proposées dans ce domaine.

Chapitre 3

Approches multicritères pour l'estimation de pertinence des documents en RI

"***Relevance*** *:*
. . . relation to the matter at hand
. . . the ability (as of an information retrieval system) to retrieve material that satisfies the needs of the user."
- Merriam-Webster Dictionary Online.

"***Agréger*** *:*
. . . Action de réunir des éléments distincts pour former un tout homogène.
- Larousse (2006)."

3.1 Introduction

Le problème de la combinaison multicritères a été abondamment étudié dans plusieurs domaines de recherche tels que la théorie du choix social (Arrow, 1974), les problèmes de prise de décision multicritères (Steuer, 1986; Roy, 1991; Bouyssou *et al.*, 2006) et la fusion de donnés (Ah-Pine, 2008; Cormack *et al.*, 2009). La combinaison multicritères s'impose généralement lorsque,

pour une tâche donnée, nous disposons d'un ensemble d'alternatives qui devraient être jugées par rapport à un nombre bien défini de critères, et nous sommes amenés à produire un score global pour chacune de ces alternatives selon chaque critère. Il faut néanmoins préciser que, selon le domaine, le terme employé n'est pas toujours "combinaison" car même si la finalité est d'unir plusieurs critères, le contexte, le type des critères et les méthodes s'avèrent très différentes. Selon le contexte de recherche, il peut s'agir de fusion de critères, agrégation de critères ou d'aide à la décision multicritères. En RI, l'agrégation multicritères a connu une attention particulière dans les deux dernières décennies. Son émergence dans la communauté de RI est liée principalement au caractère multidimensionnel du concept de pertinence. Comme nous l'avons montré au chapitre 2, de nombreux travaux (Borlund, 2003; Taylor *et al.*, 2007; Saracevic, 2007b) ont identifié plusieurs critères, au-delà de la dimension thématique, qui peuvent affecter la perception de pertinence des utilisateurs. Une direction de recherche prometteuse, qui a émergé par conséquence, consiste à considérer, à la fois, tous les critères ayant un impact sur la notion de pertinence et pouvant influencer les jugements de pertinence des utilisateurs, afin d'améliorer la précision des systèmes de RI. Le défi majeur à relever dans ce contexte consiste alors à trouver les méthodes appropriées pour pouvoir agréger les différentes dimensions ou critères afin d'aboutir à un seul score de pertinence des documents. Ceci s'est imposé de manière cruciale dans de nombreuses applications tels que la RI mobile (Göker et Myrhaug, 2008; Cong *et al.*, 2009; Bouidghaghen *et al.*, 2011b), la RI sociale (Duan *et al.*, 2010; Nagmoti *et al.*, 2010; Ben Jabeur *et al.*, 2010; Metzler et Cai, 2011; Becker *et al.*, 2011; Ounis *et al.*, 2011; Chen *et al.*, 2012) et la RI personnalisée (Sieg *et al.*, 2007; Daoud *et al.*, 2010; Bouidghaghen *et al.*, 2011a; da Costa Pereira *et al.*, 2012). Ces travaux ont mis l'accent sur l'utilité de considérer plusieurs dimensions de pertinence autres que la dimension thématique, et ont proposé différentes stratégies de combinaison de critères :

- Les moyennes arithmétique pondérées (Yager, 1988) et les stratégies de combinaison linéaire (Vogt et Cottrell, 1999; Larkey *et al.*, 2000; Si et Callan, 2002; Craswell *et al.*, 2005; Gerani *et al.*, 2012), dans lesquelles le score global est simplement une somme ou un produit linéaire pondéré des scores partiels.
- Les méthodes d'agrégation prioritaires (da Costa Pereira *et al.*, 2009, 2012) définissent des relations de priorité entre les différents critères.
- Les méthodes d'agrégation d'ordonnancements (Farah et Vanderpooten, 2007, 2008; Wei *et al.*, 2010), dans lesquelles l'aspect multidimensionnel

de la pertinence est quasiment ignoré et la combinaison ne repose, très souvent, que sur le critère thématique, en se basant sur un ensemble hétérogène d'ordonnancements de documents.
– Les approches d'apprentissage d'ordonnancements qui consistent à combiner plusieurs descripteurs en se basant sur des approches issues de l'apprentissage automatique (Liu, 2009; Cao *et al.*, 2007; Joachims, 2006).

Dans ce chapitre, nous détaillons toutes ces approches et nous présentons également d'autres méthodes issues du domaine de l'aide à la décision multicritères qui pourraient être appliquées dans le domaine de RI. Dans la section 3.2, nous commençons par une définition et une formalisation générale du problème d'agrégation. Ensuite, nous présentons une classification des approches d'agrégation multicritères. Dans la section 3.3, nous donnons un aperçu et une formalisation des approches d'agrégation basées sur les valeurs. Nous définissons le principe d'agrégation multicritères des approches les plus utilisées dans la littérature indépendamment du cadre dans lesquels elles ont été appliqués. Les méthodes d'agrégation de listes sont présentées dans la section 3.4. Dans la section 3.5, nous donnons le principe d'agrégation de pertinence multidimensionnelle et nous présentons les approches d'agrégation de valeurs et de listes qui ont été appliquées en RI. La section 3.6 conclut le chapitre.

3.2 À propos de l'agrégation multicritères

Dans cette section, nous présentons le principe d'agrégation multicritères en général et nous proposons une catégorisation des approches d'agrégation guidée par la manière par laquelle les critères sont agrégés.

3.2.1 Description du problème

Les approches multicritères en général et les opérateurs d'agrégation (ou fonctions d'agrégation) en particulier sont utilisés dans différents domaines du raisonnement humain pour l'élaboration d'une décision finale (Bouchon-Meunier et Marsala, 2003). Les fonctions d'agrégation sont généralement définies et utilisées pour combiner et résumer plusieurs valeurs, souvent numériques, en une seule, de telle sorte que le résultat final de l'agrégation prenne en compte, d'une manière prescrite, toutes les valeurs individuelles. Par exemple, supposons que plusieurs chercheurs forment des jugements

quantifiables sur l'importance d'un ensemble de publications scientifiques (qualité, originalité, nouveauté, etc) ou même sur le ratio de deux telles mesures (combien plus originale, nouvelle une publication est-elle par rapport à une autre). En agrégation multicritères, l'ensemble de publications est appelé *alternatives*. L'objectif est alors de définir une règle de décision qui permet de bâtir une relation de préférence ou de similarité (ou consensus) sur l'ensemble de ces alternatives. Pour atteindre un consensus sur les jugements, plusieurs fonctions d'agrégation classiques ont été proposées dans la littérature.

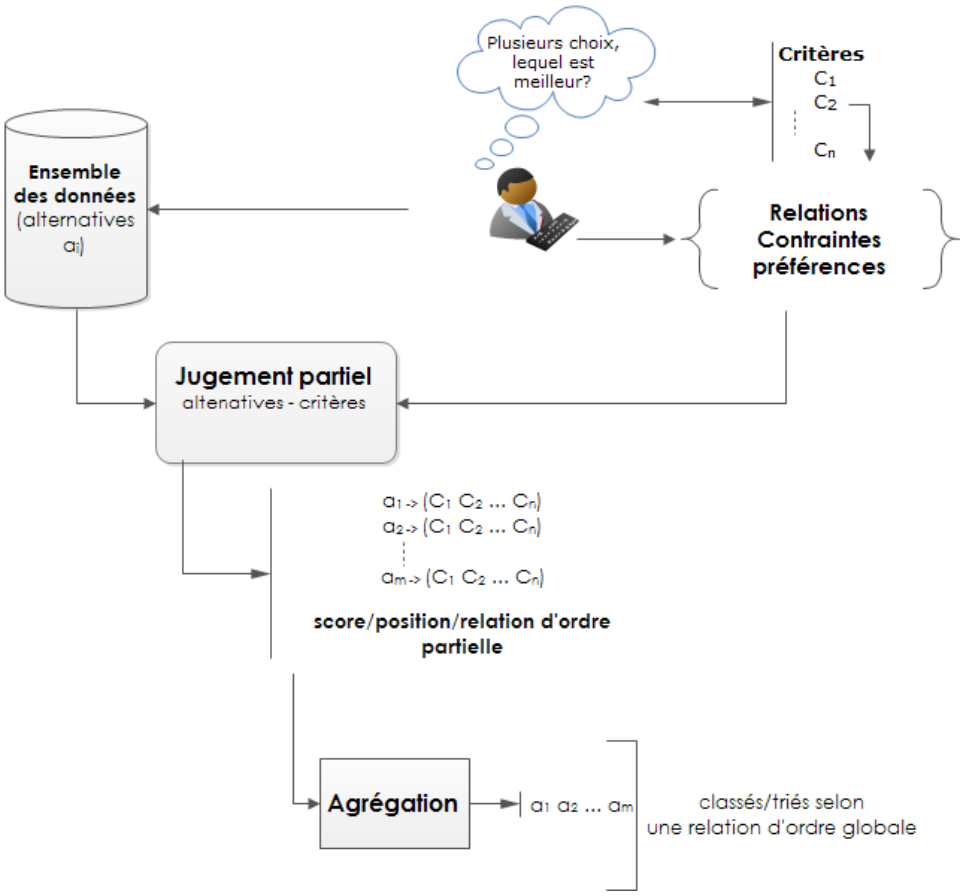


FIGURE 3.1: Architecture générale des approches d'agrégation multicritères.

Ces méthodes ont été largement exploitées dans de nombreuses disciplines bien connues comme la statistique, l'économie, la finance, l'informatique,

etc (Bouchon-Meunier et Marsala, 2003). La figure 3.1 montre les différentes étapes d'un processus d'agrégation multicritères en général. Comme le montre la figure, différentes conditions peuvent être définies sur les données en entrée, et ceci conduit également à des résultats différents selon la nature des contraintes définies. Ainsi, nous pouvons distinguer deux types d'approches d'agrégation :

– *Agréger puis comparer* : calculer une valeur globale pour chaque alternative, puis préférer celle qui obtient la meilleure valeur ;
– *Comparaison par paires* : établir la préférence entre deux alternatives en fonction des degrés de surclassement obtenus sur chaque critère.

D'une manière générale, une fonction d'agrégation peut être formalisée comme suit :

Définition 1 Fonction d'agrégation.

Soit $\mathcal{V} = \{v_1(a), \ldots, v_n(a)\}$ l'ensemble des valeurs partiels d'une alternative $a \in \mathcal{A}$, qui sont obtenus suivant un ensemble de critères $\mathcal{C} = \{c_1, \ldots, c_n\}$, la valeur de préférence globale de a est donnée par une fonction $f : \mathcal{V} \to \mathbb{R}$ obtenue suivant une agrégation de critères à l'aide d'une fonction d'agrégation $g : \mathbb{R}^n \to \mathbb{R}$:

$$f(a) = g(v_1(a), \ldots, v_n(a)) \tag{3.1}$$

Ainsi, si on considère deux alternatives a et $a' \in \mathcal{A}$, on peut définir une relation de préférence globale $\preceq_f$ qui permet de favoriser l'une des deux alternatives sur l'ensemble global des critères. Dans les approches de type "agréger puis comparer", la préférence *a est mieux jugée que a'* est représentée par : $a \preceq_f a' \Leftrightarrow g(v_1(a), \ldots, v_n(a)) \geq g(v_1(a'), \ldots, v_n(a'))$.

Dans les approches de type "comparaison par paires", les relations de préférences sont représentées au niveau des critères (niveau local ou partiel). Donc, les alternatives sont jugées selon un critère donné et non sur la totalité de l'ensemble $\mathcal{C}$.

3.2.2 Classification des approches

Les approches multicritères peuvent être classées selon le type de l'agrégation (score, position, préférence, etc). Comme montré dans la figure 3.2, nous catégorisons ces approches en deux classes principales :

1. *Agrégation de valeurs :* où toutes les alternatives sont données avec

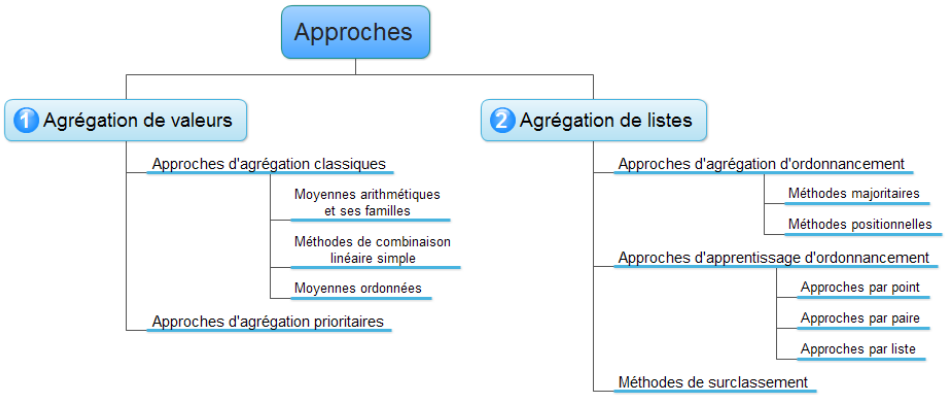


FIGURE 3.2: Classification des approches d'agrégation multicritères.

des scores partiels avec tous les critères considérés. L'agrégation dans ce cas consiste alors à combiner des valeurs numériques. Cette catégorie d'approches d'agrégation inclut les méthodes de combinaison linéaire ainsi que divers opérateurs d'agrégation prioritaires et d'autres méthodes classiques issues du domaine d'aide à la prise de décision multicritères.

2. *Agrégation de listes :* où nous disposons des listes d'ordonnancements plutôt que des critères. Bien que ces listes peuvent être assimilées à des critères, le problème reste un peu différent dans la mesure où nous pouvons être confrontés à agréger des positions et non plus des valeurs. Ces positions représentent le classement des alternatives dans les listes. Il est important de noter que dans certaines situations, nous pouvons avoir des alternatives qui n'ont pas de positions ou de scores dans certaines listes, contrairement au problème d'agrégation de valeurs.

Les approches d'agrégation présentés dans la figure 3.2 peuvent être aussi classées en plusieurs autres catégories ; compensatoires, non compensatoires, conjonctives, disjonctives ou basés sur la pondération (Hwang et Yoon, 1981). Les méthodes compensatoires sont basées sur l'hypothèse qu'un score faible d'une alternative donnée qui est obtenu suivant un critère important (préféré) pourrait être compensé par un score élevé d'un autre critère important. Les opérateurs compensatoires sont généralement compris entre un maximum et un minimum, i.e., ils ne sont ni conjonctifs ni disjonctifs. La moyenne pondérée est la fonction la plus représentative de cette famille d'opérateurs (Kolmogorov, 1930; Aczel, 1948). Une autre famille d'opéra-

teurs amplement étudiée dans littérature concerne les moyennes ordonnées (Yager, 1988). Ces opérateurs sont à l'intersection des deux opérateurs "min" et "max" classiques. Ainsi, les opérateurs *t-norm* et *t-conorm* (Schweizer et Sklar, 1960, 1983; Menger, 1942) on été introduit pour généraliser les deux opérateurs conjonctifs ("Et") disjonctifs ("Ou"). Les opérateurs d'agrégation compensatoires requirent souvent un ensemble de priorités ou de préférences. Ces derniers sont généralement exprimés à l'aide d'un ensemble de poids ou de fonctions de priorité sur les critères.

De l'autre côté, les fonctions d'agrégation non compensatoires telles que "min" ou "max" (Fox et Shaw, 1993) sont généralement dominées par un seul critère, i.e., le meilleur ou le plus faible score. Un des inconvénients de cette famille d'opérateurs est qu'une grande partie des scores sont ignorés dans le processus final d'agrégation.

D'autres classifications sont également possibles, se basant sur le fait que la méthode utilisée est supervisée ou non, ou aussi selon le cadre d'application. Nous tenons à classer les approches selon la méthode avec laquelle les scores (ou les positions) sont agrégées. Dans ce chapitre, nous adoptons la classification basée sur le type de combinaison (valeurs ou listes), car elle est la plus proche du domaine de RI.

3.3 Approches d'agrégation de valeurs

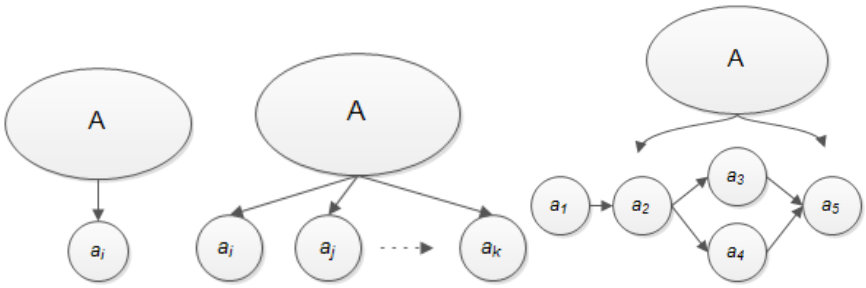
L'objectif de cette section est de présenter les opérateurs basés sur l'agrégation de valeurs. Nous commençons, tout d'abord, par expliquer le principe des méthodes d'agrégation classiques. Nous verrons ensuite les approches d'agrégation prioritaires existants dans la littérature.

3.3.1 Description du problème

Les approches d'agrégation de valeurs trouvent leur origine dans le domaine d'aide de prise à la décision. Le processus d'aide à la décision consiste généralement à supporter un utilisateur dans le processus de prise de décision dans le cas où il est confronté à des problèmes complexes, par exemple, de planification, de choix ou d'évaluation, etc. Étant donnés que la "réalité" est multidimensionnelle et qu'il existe toujours plusieurs critères à satisfaire, ces systèmes se sont orientés vers une approche multicritères. L'objectif d'une aide multicritères à la décision est donc de proposer à un utilisateur les

meilleures alternatives répondant le plus à ses besoins (et qui soit la plus optimale à son égard), étant donné un ensemble de critères et contraintes qui doivent être pris en compte dans le processus de décision. Si les critères sont partiellement conflictuels, ou si plusieurs systèmes devraient juger l'ensemble des alternatives, la prise de décision devient difficile. D'autant plus que la notion d'optimisation "dans l'absolu" est vide de sens en décision multicritères, car il n'existe généralement pas d'alternative optimisant tous les critères simultanément. Il est donc nécessaire de prendre en compte de l'information supplémentaire, en particulier l'importance relative de chaque critère et les compensations possibles entre les différents critères. Chacune de ces difficultés a donné naissance à un champ particulier de recherche (décision dans l'incertain, décision multicritères, etc). (Roy, 2003) distingue trois types de problématiques en aide à la décision, illustré dans la figure 3.3 :

– La problématique du choix, qui consiste à vouloir choisir la ou les solutions considérées comme optimales pour le problème considéré ;
– la problématique du classement (*ranking*), qui consiste à vouloir classer du premier au dernier toutes les solutions connues du problème considéré ;
– la problématique du tri (*sorting*), qui consiste à affecter les solutions à des catégories (ordonnées ou non).



La problématique du choix — la problématique du classement — la problématique du tri

FIGURE 3.3: Problématiques liées aux méthodes multicritères.

Dans notre travail, nous nous intéressons plus particulièrement aux méthodes qui mettent l'accent sur problématique du tri. Formellement, dans un problème d'AMD, comme nous l'avons déjà cité, on dispose souvent d'un ensemble d'alternatives (ou objets) $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ et l'objectif est de

choisir les meilleures parmi ces alternatives en présence d'un ensemble de critères $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ et un ensemble de préférences. Le point de départ consiste à formuler l'ensemble de préférences et évaluer l'impact de chaque alternative selon chaque critère (Jankowski, 1995). Cet impact est appelé selon le domaine d'application, où une évaluation de performance qui est définie selon une relation de préférence partielle $\preceq_{c_i}$ $(1 \leq i \leq n)$. Ensuite, ces préférences peuvent être formulées, comme c'est le cas pour la moyenne arithmétique pondérée, comme un vecteur normalisé de poids $W = (w_l, w_2, \ldots, w_n)$ (où $0 \leq w_i \leq 1$). Chaque poids w_i représente le degré d'importance du critère c_i. Par exemple, pour évaluer l'importance globale de l'alternative a_1, un ensemble de scores partiels$(v_1(a_1), \ldots, v_n(a_1))$ est tout d'abord calculé sur tous les critères, puis agrégé en se basant les poids de chacun. Le résultat final est alors un ensemble d'alternatives triées ou classées selon l'ensemble de préférences préalablement défini. Nous développons dans la suite, les différentes méthodologies d'agrégation issues du domaine d'AMD se basant sur les scores.

3.3.2 Approches d'agrégation classiques

Dans le domaine de RI, pour pouvoir ordonner les documents dans un ordre décroissant de pertinence, il est souvent nécessaire de combiner différentes sources de pertinence. La plupart des modèles calculent pour chaque document un score partiel, appelé *rsv* ("*retrieval status value*"), selon chacun des critères, et le combine avec les scores des autres critères. Les documents sont finalement ordonnés dans l'ordre décroissant du score global obtenu. Dans cette section, nous traitons les méthodes classiques de combinaison multi-critères de pertinence. Nous donnons une formalisation de chacune des ces méthodes et nous proposons une classification basée sur la technique avec laquelle les scores sont combinés.

3.3.2.1 Moyennes arithmétiques et mécanismes de combinaison linéaire classiques : Principes

Le concept de moyenne a donné lieu aujourd'hui à un champs d'étude très vaste avec une variété impressionnante d'applications. En fait, une abondante littérature sur les propriétés de plusieurs moyennes (tels que la moyenne arithmétique, géométrique, etc.) a été déjà introduite et continue à se développer aujourd'hui (Bouyssou *et al.*, 2006). Le concept de moyenne

a été initialement défini par (Chisini, 1929) comme suit.

Définition 2 Moyenne..

Soit $f = g(x_1, \ldots, x_n)$ une fonction de n variables indépendantes $x_1, \ldots, x_n$ représentant des quantités homogènes. Une moyenne de $x_1, \ldots, x_n$ par rapport à la fonction g est un nombre M tel que, si tous les x_i sont remplacés par M, la valeur de la fonction reste inchangée, c'est-à-dire,

$$g(M, \ldots, M) = g(x_1, \ldots, x_n) \tag{3.2}$$

Lorsque g est la somme, le produit, la somme des carrés, la somme des inverses, ou encore la somme des exponentielles, la solution de l'équation de Chisini correspond respectivement à la moyenne arithmétique, la moyenne géométrique, la moyenne quadratique, la moyenne harmonique, et la moyenne exponentielle. Le tableau 3.1 illustre les formules de ces méthodes.

Type de moyenne	$f(x)$	$M^{(n)(x_1,\ldots,x_n)}$
Arithmétique	x	$\frac{1}{n}\sum_{i=1}^{n} x_i$
Géométrique	$log(x)$	$(\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$
Quadratique	x^2	$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$
Harmonique	x^{-1}	$\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$
Exponentielle	$\exp^{\alpha x}(\alpha \in \mathcal{R}^*)$	$\frac{1}{\alpha} ln\left(\frac{1}{n}\sum_{i=1}^{n}\exp^{\alpha x_i}\right)$

TABLE 3.1: Exemples de moyennes arithmétiques.

3.3.2.2 Moyennes ordonnées

Une autre famille d'opérateurs d'agrégation qui a été étudié dans la littérature est celle des opérateurs se basant sur la somme pondérée ordonnée (communément appelées OWA pour *Ordered Weighted Average*). Cet opérateur a été initialement introduit par (Yager, 1988). L'idée derrière cet opérateur d'agrégation est d'essayer de relier deux cas extrêmes : "tous les critères doivent être satisfaits" (comme pour une opération AND classique), et un "seul critère satisfait suffit" (OR). Donc, pour ce faire, il suffit de récupérer pour chaque ensemble d'alternatives l'ensemble des scores partiels

$(v_i(a))$ (performances ou aussi utilités), puis de le trier en ordre décroissant. Ensuite, il suffit d'effectuer une somme pondérée de ces scores par des coefficients prédéfinis. L'opérateur OWA est défini par :

$$OWA(v_1(a), \ldots, v_n(a)) = \sum_{i=1}^{n} w_i \cdot a_{(i)} \tag{3.3}$$

Avec $W = (w_l, w_2, \ldots, w_n)$ un vecteur de poids normalisé, $w_i \in [0, 1]$ tel que $\sum_{i=1}^{n} w_i = 1$ et où la notation $(\cdot)$ indique une permutation des indices telle que $a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(n)}$. Ainsi, le poids n'est plus sur les scores partiels mais sur les rangs. Cet opérateur possède en outre la particularité de généraliser un certain nombre d'opérateurs que nous avons déjà mentionné jusqu'ici : minimum, maximum et moyenne arithmétique :

- Si $w_1 = 1$ (et donc $w_i = 0$ pour $i > 1$), OWA généralise l'opérateur "min" ;
- Si $w_n = n$, c'est l'opérateur "max" ;
- n est impair, $w_{\frac{n+1}{2}} = 1$, c'est la médiane. Si n est pair, la médiane est défini par $w_{\frac{n}{2}} = w_{\frac{n}{2}+1} = \frac{1}{2}$.

3.3.3 Approches d'agrégation prioritaires

Comme pour les approches d'agrégation classiques, l'agrégation prioritaire peut aussi être formalisée comme un problème de prise de décision multicritères (Cf. Section 3.3.1), dans lequel nous avons :

- L'ensemble des alternatives $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$
- L'ensemble des critères $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$
- Des préférences sur les différents critères qui sont représentées par des relations de priorité $\preceq_{c_i}$; $c_1 \preceq c_2$ est interprété par c_1 est plus prioritaire que c_2
- La fonction d'agrégation $g(v_1(a), \ldots, v_n(a))$ calculant le score global d'une alternative a par rapport à l'ensemble total des critères. Le score $v_i(a)$ représente le degré de satisfaction d'une alternative a par rapport au critère c_i.

Pour calculer le score global, un opérateur d'agrégation prioritaire affecte des poids d'importance à chaque critère de l'ensemble $\mathcal{C}$ de manière à ce que les critères les plus prioritaires aient les poids les plus élevés. Il est à noter aussi qu'un poids d'importance d'un critère c_i dépend aussi des degrés de satisfaction des alternatives et des poids d'importance des critères les plus prioritaires (que c_i). Pour simplifier les notations, on dénote par c_1 le critère

le plus préféré et c_n le moins prioritaire. Donc on assume que c_i est plus préféré que c_j ($c_i \preceq c_j$) si et seulement si $i < j$. Cette intuition peut être formalisée comme suit :

- Pour chaque alternative a, le poids du critère le plus prioritaire c_1 est 1, i.e., par définition, $\forall d,\ \lambda_1 = 1$;
- Le poids des autres critères c_i ($i \in [2..n]$) est calculé de la façon suivante :

$$\lambda_i = \lambda_{i-1} \cdot v_{i-1}(a) \tag{3.4}$$

  Où $v_{i-1}(a)$ est le degré de satisfaction du critère c_{i-1} par l'alternative a, et λ_{i-1} ($\in [0..1]$) est le poids d'importance de c_{i-1}.

L'intuition derrière l'idée de priorité entre les critères vient du fait que dans certaines applications réelles, l'importance d'un critère pourrait être dépendante de la satisfaction d'un autre critère plus prioritaire.

3.4 Approches d'agrégation de listes

3.4.1 Approches d'agrégation d'ordonnancements

L'agrégation d'ordonnancements, appelée aussi fusion d'ordonnancements, consiste à agréger un certain nombre de listes partielles d'ordonnancements (le nombre de moteurs de recherches) de manière à optimiser la performance de la combinaison (Dwork *et al.*, 2001; Wei *et al.*, 2010). Dans ce contexte, les approches existantes dans la communauté de RI peuvent être classées selon deux façons ; soit par *(i)* le type d'apprentissage requis : supervisé, semi-supervisé et non supervisé (Wei *et al.*, 2010), soit *(ii)* selon le type d'agrégation et les données requises pour ce processus (Riker, 1982; Farah et Vanderpooten, 2007) où nous distinguons deux familles de méthodes : les méthodes d'agrégation *positionnelles* et les méthodes d'agrégation *majoritaires*. Dans cette section, nous allons adopter la deuxième classification.

3.4.1.1 Principe de l'agrégation d'ordonnancements

L'agrégation d'ordonnancements est un problème qui a été largement étudié en plusieurs domaines d'applications, par exemple, dans les méta-recherche (Aslam et Montague, 2001; Akritidis *et al.*, 2011), la fusion d'images et plusieurs d'autres domaines. L'objectif est de trouver un consensus sur un

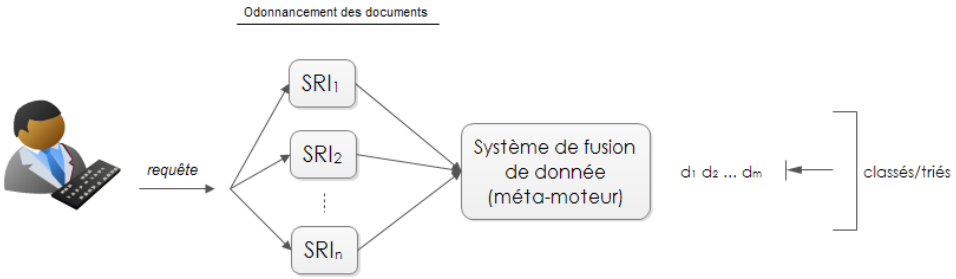


FIGURE 3.4: Principe de fusion d'agrégation des méthodes de fusion d'ordonnancements (méta-moteurs).

ensemble de documents étant donné un ensemble d'ordonnancements de plusieurs systèmes (Renda et Straccia, 2003). En méta-recherche, les systèmes représentent généralement des moteurs de recherche et les alternatives sont les documents retournés suivant une requête donnée. Le problème consiste alors à combiner les listes de sorte à ce que la combinaison et la liste finale de documents soit la plus optimale possible (Aslam et Montague, 2001). La figure 3.4 montre le fonctionnement des méta-moteurs pour fusionner ces résultats.

Comme souligné par Riker (1982), les méthodes émanant de ce domaine peuvent être classées en deux catégories, selon s'ils sont basées sur les ordres (positions) ou les scores des documents (*i.e.,* suivant : si les systèmes SRI_k retournent des scores ou des positions). Les méthodes basées sur les ordres sont appelées "méthodes positionnelles", et celles basées sur les scores sont appelées "méthodes majoritaires". Ces deux familles de méthodes d'agrégation ont retrouvé leurs origines dans les travaux pionniers de Borda (Borda, 1781) et Condorcet Condorcet (1785), respectivement, dans la littérature du choix social.

3.4.1.2 Méthodes majoritaires

Ces méthodes utilisent des comparaisons paire à paire entre les documents ou les items des ordonnancements retournés, et sont principalement basées sur l'agrégation des relations d'ordre en utilisant des critères d'association tels que la procédure Condorcet (Condorcet, 1785) ou la distance de *Kendall* (Fagin *et al.*, 2003).

La méthode de Condorcet est une méthode de vote très populaire, et qui a

été parmi les premières méthodes d'agrégation d'ordonnancements utilisées dans la littérature. Le principe de cette méthode est que l'unique vainqueur est celui, s'il existe, qui, comparé tour à tour à tous les autres candidats, s'avérerait à chaque fois être le candidat préféré. Considérons par exemple, une situation de vote, où 90 votants élisent, à la majorité relative, un candidat parmi a, b et c :

Nombre de votants	*Classement*
34 votants	$c\ b\ a$
29 votants	$a\ b\ c$
27 votants	$b\ a\ c$

Dans cet exemple, la règle majoritaire désigne le candidat c (34 suffrages, contre 29 pour a et 27 pour b). Et pourtant, une majorité de 56 votants le place en dernier. Toutefois, la méthode de Condorcet, dont le principe est : "*un candidat est élu s'il bat tous les autres à la majorité simple (dans un duel par paires)*" désigne b comme vainqueur ($34+27$ voix contre a et $29+27$ voix contre c). Étant donnée qu'il existe des cas où il n'y pas toujours un vainqueur (e.g., cas où on a une relation cyclique : a bat b, b bat c et c bat a), des variations de la méthodes Condorcet, comme la famille ELECTRE ont été proposées dans le domaine d'aide à la décision multicritères. Ces méthodes seront discutées dans la section 3.4.2.

D'autres approches se basant sur les chaînes de Markov (CM) ont aussi montré leur efficacité en combinant divers listes d'ordonnancements (Dwork *et al.*, 2001). Dwork *et al.* (2001) ont proposé un modèle où les états correspondent aux documents à classer et les transitions d'un état à un autre varient selon le classement des documents, *i.e.,* pour la majorité des ordonnancements, un document donné est mieux classé qu'un autre. Dans le même contexte, les auteurs ont proposé 4 chaînes de Markov spécifiques, et l'évaluation expérimentale a montré que la meilleure est la suivante (cf., Renda et Straccia (2003)) :

– CM4 : se déplacer de l'état actuel d_i à l'état prochain $d_{i'}$ en choisissant, dans un premier temps, de façon uniforme un document $d_{i'}$ à partir de la collection D. Si, pour la majorité des ordonnancements, nous avons $pos_j(d_{i'}) \leq pos_j(d_i)$, alors se déplacer à $d_{i'}$, sinon rester à l'état d_i. $pos_{r_j}(d_i)$ est la position de d_i dans l'ordonnancement r_j.

La distribution stationnaire de la chaîne CM4 est utilisée pour classer les documents.

3.4.1.3 Méthodes positionnelles

Dans ce cadre, il s'agit d'attribuer des scores aux documents à classer selon les *classements* (*ranks*) qu'ils reçoivent, puis d'agréger ces scores en utilisant des techniques différentes.

La méthode de *Borda* est la première méthode proposée pour ce cadre d'agrégation (Borda, 1781). La technique utilisée consiste à assigner les scores en se basant sur les positions des documents dans les listes d'ordonnancements. Cette méthode est souvent adaptée aux problèmes de classification par fusion de classifieurs, où les classes sont considérés, si l'on utilise la terminologie de la littérature de vote, comme des votants et les classes comme des candidats. Chaque candidat reçoit les points de la part de chaque votant selon sa position dans la liste ordonnée. Ainsi, le candidat préféré reçoit le plus grand nombre de vote. Les candidats sont ainsi ordonnés suivant le score donné par la formule suivante :

$$score(ca) = \sum_{j=1}^{n} pos_{r_j}(d_i) \tag{3.5}$$

Où n est le nombre d'ordonnancements et $pos_{r_j}(d_i)$ est la position de d_i dans r_j.

Plus tard, Fox et Shaw (1993) ont proposé plusieurs stratégies de combinaison tels que les opérateurs *combSUM*, *combMIN*, *combMAX*, *combANZ* et *combMNZ*. Les trois premières méthodes correspondent aux opérateurs somme, min et max, respectivement. *combANZ* et *combMNZ* divisent et multiplient le score fourni par *combSUM* par la position des candidats.

Marden (1995) a proposé une nouvelle méthode pour l'optimisation d'ordonnancements qui minimise la distance *footrule* de Spearman. Étant données deux listes r_j et $r_{j'}$, la distance est donnée par :

$$F(r_j, r_{j'}) = \sum_{i=1}^{n} |pos_{r_j}(d_i) - pos_{r_{j'}}(d_i)| \tag{3.6}$$

3.4.2 Approches de surclassement

Contrairement aux approches d'agrégation multicritères précédemment présentées, les approches de surclassement (Roy *et al.*, 1978) visent à construire

des relations de dominance (ou surclassement) entre les éléments de l'ensemble A des alternatives. Une relation de surclassement est une relation binaire S définie dans l'ensemble A telle que : $a_i S a_j$ s'il y a suffisamment d'arguments pour admettre que a_i est au moins aussi bonne que a_j, étant données les évaluations ou les scores de ces deux alternatives selon les critères en entrée. La relation S est généralement construite à l'aide des comparaisons entres toutes les paires d'alternatives de A. Il existe de nombreuses méthodes de surclassement qui ont été proposées dans la littérature, parmi lesquelles on cite la famille des méthodes Electre (Roy, 1991) qui sont les plus connues du domaine de décision multicritères, PROMETHE (Brans et Vincke, 1985; Brans *et al.*, 1984), TACTIC (Vansnick, 1986), GAIA, etc. Tant dis que la plupart de ces méthodes s'intéressent à des problématiques de choix ou de classement, la méthodes Electre-Tri est la méthode la plus proche qui traite la problématique de tri. Dans Eletectre-Tri, l'importance des critères dans la prise de décision est évaluée grâce à des relations de concordance et discordance entre les différentes paires d'alternatives.

3.4.3 Approches d'apprentissage d'ordonnancements

Dans cette section, nous introduisons les approches d'apprentissage d'ordonnancements (ou *learning to rank*).

3.4.3.1 Description du problème

L'apprentissage d'ordonnancements est un sous domaine majeur de l'apprentissage automatique (Liu, 2009). L'objectif principal des algorithmes proposés dans ce domaine consiste à combiner plusieurs descripteurs de pertinence afin d'optimiser l'ordonnancement des documents et ce en se basant sur des approches issues de l'apprentissage automatique. Ces techniques sont souvent appliquées par les moteurs de recherche pour la combinaison de différents modèles de pondération de documents ou autres descripteurs liés à la requête (Liu, 2009). Les modèles d'apprentissage d'ordonnancements génèrent un modèle qui pourrait représenter au mieux la fonction d'ordonnancements. La forme du modèle généré diffère selon la technique utilisée, pour certaines elle peut être représentée par un vecteur de poids pour combiner de façon linéaire chaque descripteur (Metzler, 2007; Xu et Li, 2007), pour d'autres elle peut représenter un réseau de neurones (Burges *et al.*, 2005) ou une série d'arbres de décision (Weinberger *et al.*, 2010). Avant de procéder à

la génération du modèle, un algorithme d'apprentissage d'ordonnancements commence par la représentation des couples de requêtes-document dans l'espace des descripteurs (features). Considérons une requête q, un document d_j et d le nombre de descripteurs, alors le couple requêtes-document (q, d_j) est représenté par le vecteur $x = \phi(q, d_j) \in \mathcal{R}^d$ où ϕ est un extracteur de descripteur tel que $x_i = \phi_i(q, d_j)$ est la valeur du descripteur i pour le couple (q, d_j). Les descripteurs de pertinence communément utilisés incluent souvent des mesures de similarité tel que BM25, le degré d'importance PageRank ou d'autres caractéristiques du document ou de la requête.

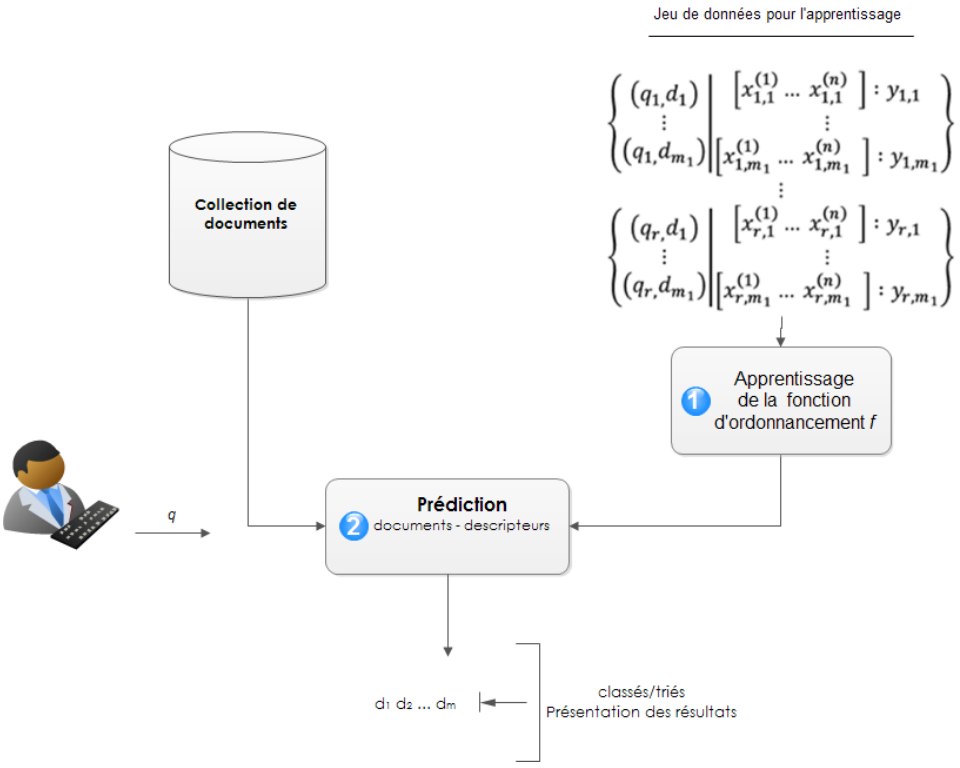


FIGURE 3.5: Schéma général des approches d'apprentissage d'ordonnancements.

Dans un modèle d'apprentissage d'ordonnancements, le processus général de classement des documents se décompose en deux étapes principales, comme montré dans la figure 3.5 (Liu, 2009) :

– *Apprentissage de la fonction d'ordonnancements* : dans cette étape, le mo-

dèle prend en entrée un jeu de données comprenant l'ensemble des paires requête-documents (q_i, d_j), où chaque paire est représentée par le vecteur des descripteurs $x_{i,j} \in \mathcal{R}^d$, tel que $x_{i,j} = x_{i,j}^{(1)} \dots x_{i,j}^{(d)}$. A chaque paire ($q_i$, d_j) est associé un score de pertinence $y_{i,j}$ mesurant la correspondance entre la requête et le document. Ce score peut être soit un nombre réel, soit un entier représentant le degré ou la classe de pertinence du document (e.g., 0 pour les non pertinents, 2 pour un très pertinent, etc). Ces scores (ou *labels*) sont généralement donnés manuellement par des assesseurs, et utilisés pour apprendre la fonction d'ordonnancements. Cette fonction permet alors de prédire les scores de pertinence des documents à travers la minimisation d'une fonction *objectif* (*loss function*) (i.e., avoir la plus petite erreur possible).

– *Ordonnancement des documents* : l'algorithme utilise la fonction apprise dans la première étape pour la prédiction de la pertinence des nouveaux documents (i.e., n'ayant pas fait partie de l'apprentissage) suivant chaque requête. Le modèle permet alors de générer pour chaque requête, l'ensemble des documents ordonnés selon les valeurs données par la fonction objectif, avec pour chacun un score ou un degré de pertinence.

Au cours de la dernière décennie, un grand nombre d'algorithmes ont été proposés pour l'apprentissage d'ordonnancements. Ils sont généralement regroupés sous trois grands types d'approches : par point (*pointwise*), par paire (*pairwise*) et par liste (*listwise*) (Liu, 2009). Ces approches sont détaillées dans la suite.

3.4.3.2 Méthodes par point (*pointwise*)

Dans les approches par point, la fonction *objectif* est définie sur des objets uniques. Comme nous l'avons déjà mentionné, les jugements de pertinence peuvent être soit des scores réels soit des degrés de pertinence ou même non ordonnés (pertinent/non pertinent).

Dans le cas où les scores sont des nombres réels, le problème d'agrégation d'ordonnancements peut être ramené à un problème de régression linéaire. Pour chaque document d_j, l'algorithme apprend une fonction $f(y_j, \overline{y_j}) = (y_j - \overline{y_j})^2$ qui minimise l'écart entre y_j le score de pertinence de référence et $\overline{y_j}$ le score de pertinence prédit, traduisant l'écart entre la valeur prédite et la valeur attendue. Dans le cas où les scores sont binaires le problème d'ordonnancements peut être ramené à un problème de discrimination. Dans le cas où les scores de pertinence sont des variables ordonnées, on exploite gé-

néralement des méthodes de régression ordinale, qui permettent de prendre en compte l'ordre relatif entre les classes pour apprendre le modèle (Liu, 2009). L'approche d'ordonnancements par point est la plus simple à mettre en œuvre, mais aussi la moins performante, car elle ne prend pas en compte l'ordre relatif des documents, contrairement à l'approche par paire, qui lui est généralement préférée.

3.4.3.3 Méthodes par paire (*pairwise*)

Le principe d'ordonnancements consiste ici à faire des comparaisons entre les paires des documents, à travers des préférences, pour déterminer lequel est plus pertinent. L'objectif est donc d'apprendre la fonction qui permet de discriminer au mieux les paires de documents et de leur affecter la classe correspondante. Parmi les algorithmes les plus connus abordant ce type de problème, les SVM (Joachims, 2006), les réseaux de neurones (Burges *et al.*, 2005) et les arbres de décision.

Le principe de l'algorithme RankSVM est de rechercher l'hyperplan qui sépare de façon optimale les documents non pertinents des documents pertinents dans l'espace des descripteurs de pertinence. Formellement, si on considère une paire de documents (d_i, d_j) associée à une requête q et représentée par le vecteur $x_q = x_i - x_j \in \mathcal{R}^d$ dans l'espace des descripteurs de pertinence, SVM l'apprentissage de la fonction objectif s'effectue à travers la résolution du problème d'optimisation suivant :

$$min_w \frac{1}{2} \|w\|_2^2 + C \sum_{q=1}^{r} l(w^T x_q) \tag{3.7}$$

où w est le vecteur des poids représentant le modèle linéaire appris, w^T son transposé, C est un paramètre permettant le contrôle des erreurs de prédiction et $l()$ est une fonction de perte telle que $l(w^T x_q) = max(0.1 - w^T x_q)$ dans RankSVM.

3.4.3.4 Méthodes par liste (*listwise*)

En ce qui concerne les approches par liste (Cao *et al.*, 2007), on considère la totalité de la liste ordonnée des documents pour chaque requête comme une instance pour l'apprentissage, à la différence de l'approche par paire

où l'on considère des comparaisons paire à paire. Par conséquence, ces approches sont capables de différencier les documents des différentes requêtes, et de considérer leur rang lors de l'apprentissage. Ces méthodes peuvent être classées en deux catégories suivant les fonctions de perte utilisées : *(i)* les méthodes considérant des fonctions de perte définies à partir de mesures de RI, et *(ii)* celles utilisant des fonctions de perte indépendantes des mesures de RI.

Parmi les algorithmes se basant sur des mesures de RI, nous trouvons AdaRank (Xu et Li, 2007), basé sur l'algorithme AdaBoost et permettant ainsi l'optimisation des métriques MAP et du NDCG respectivement. Parmi les algorithmes exploitant des fonctions objectifs indépendantes des mesure de RI, nous citons ListNet (Cao *et al.*, 2007). L'objectif de cet algorithme est de construire une fonction de perte qui mesure le nombre de permutations entre la liste de référence et la liste apprise.

3.5 Agrégation de pertinence multidimensionnelle en RI

Tandis que les recherches antérieures sur la pertinence ont été axées sur ce concept du point de vue thématique, les recherches les plus récentes ont mis l'accent sur plusieurs dimensions de pertinence différentes : représentation multidimensionnelle côté utilisateur (profil, centres d'intérêts, expertise, autorité), environnement de recherche (dispositifs utilisés, localisation géographique) et aspect temporel (Cooper, 1973; Barry, 1994; Cosijn et Ingwersen, 2000; Borlund, 2003; da Costa Pereira *et al.*, 2012; Taylor, 2012; Gerani *et al.*, 2012). Les principales constations qui ressortent des ces études sont :

- Les critères de pertinence ne sont pas indépendants les uns des autres, et généralement ceux liés au contenu, qui incluent la dimension thématique se sont données les poids d'importances les plus élevés. De plus, ces dimensions interagissent avec les autres critères (Saracevic, 2007b; Eickhoff *et al.*, 2013a) ;
- Un nombre fini et peu limité de critères est conjointement considéré par les utilisateurs pour juger la pertinence (Eickhoff *et al.*, 2013a) ;
- L'importance des critères dépend de la tâche de recherche et de la classe d'utilisateurs finaux (da Costa Pereira *et al.*, 2012).

En RI, l'agrégation des critères revient à combiner les dimensions de pertinence identifiées dans un cadre de recherche bien défini. Suivant le type

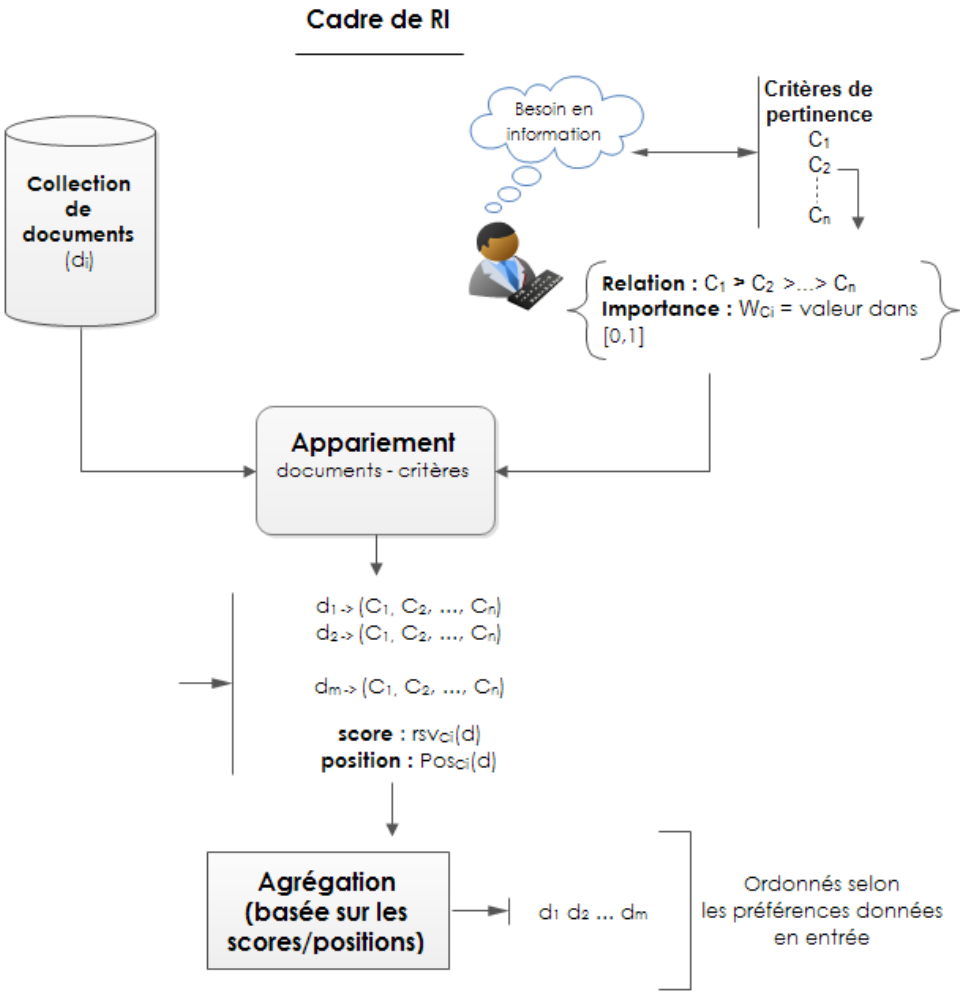


FIGURE 3.6: Instanciation de l'agrégation multicritères dans pour la combinaison de pertinence multidimensionnelle en RI.

des critères, leur propriétés et les relations définies entre eux, plusieurs méthodes d'agrégation ont été proposés, dont les plus utilisées sont la moyenne arithmétique, la médiane et bien d'autres encore. Ainsi, plusieurs familles de méthodes de combinaison multicritères ont vu leur application dans de nombreuses applications de RI telles que la RI mobile (Göker et Myrhaug, 2008; Cong *et al.*, 2009; Bouidghaghen *et al.*, 2011b), la RI sociale (Duan *et al.*, 2010; Nagmoti *et al.*, 2010; Ben Jabeur *et al.*, 2010; Metzler et Cai, 2011;

Becker *et al.*, 2011; Ounis *et al.*, 2011; Chen *et al.*, 2012) et la RI personnalisée (Sieg *et al.*, 2007; Gauch *et al.*, 2003; Daoud *et al.*, 2010; Bouidghaghen *et al.*, 2011a; da Costa Pereira *et al.*, 2012). Chaque méthode d'agrégation a sa spécificité en fonction des caractéristiques des données en entrée à ces approches. Dans le domaine de l'aide à la décision multicritères, les valeurs à agréger sont généralement des préférences (d'une alternative par rapport à une autre) ou des degrés de satisfaction (d'une alternative) relatifs à des critères. Le même principe est utilisé dans les méthodes d'agrégation prioritaires où une relation de priorité est définie entre les différents critères sur la base d'un mode de calcul de poids associés qui favorise la satisfaction du critère par rapport à un autre. Par ailleurs, dans le domaine d'aide à la décision face à l'incertain (floue), les valeurs à agréger représentent les conséquences d'une action relatives à des états de la nature. La figure 3.6 montre l'architecture générale d'une approche d'agrégation de pertinence en RI.

Dans ce qui suit, nous allons donner un aperçu des travaux sur l'agrégation multicritères en RI suivant le schéma et la classification donnés dans la section 3.2.2.

3.5.1 Approches basées sur l'agrégation de valeurs

3.5.1.1 Moyennes arithmétiques et mécanismes de combinaison linéaire classiques

Les recherches impliquant la combinaison de plusieurs sources d'évidence sont souvent basées sur des fonctions de combinaison linéaire en vue de leur simplicité et leur efficacité relative, quelque soit le domaine d'application considéré. C'est également le cas dans le domaine de RI, étant donnée la nature multidimensionnelle de la pertinence. Cette propriété a été exploitée dans différents cadres de RI, et chaque contexte représente la pertinence suivant les descripteurs ou facteurs existants qui peuvent impacter le jugement de pertinence des documents.

Par exemple, dans une tâche de RI sociale, et plus particulièrement dans Twitter, une variété de critères tels que la thématique des tweets, l'autorité des utilisateurs et beaucoup d'autres facteurs ont été exploités pour estimer la pertinence des tweets (Duan *et al.*, 2010; Nagmoti *et al.*, 2010; Ben Jabeur *et al.*, 2010; Metzler et Cai, 2011; Becker *et al.*, 2011; Ounis *et al.*, 2011; Chen *et al.*, 2012). Plusieurs de ces travaux ont été proposés dans le

cadre de la tâche Microblog de TREC, ces derniers ont exploité les signaux sociaux identifiés dans les collections de tweets fournies par la tâche pour l'amélioration de leurs systèmes de recherche (Metzler et Cai, 2011; Ounis *et al.*, 2012). Pour agréger les différents critères de pertinence identifiés, la plupart des méthodes proposées se basent sur des mécanismes de combinaisons linéaires.

Zhao *et al.* (2011) ont proposé l'utilisation de trois facteurs de pertinence existants : *TweetRank (TR)*, *FollowerRank (FR)* et *LengthRank (LR)* pour l'ordonnancement des tweets. Ensuite, ils ont raffiné le critère présence d'*URL* en calculant la fréquence (*URLRank*) plutôt que la présence, et ont ajouté à ceci deux autres critères à savoir, le nombre de *retweets* (*RetweetRank*) et le nombre de mentions (*MentionRank*). Ces critère ont été intégrés dans un modèle de combinaison linéaire qui a la forme suivante :

$$f_{MFR}(t,q) = f_{FR}(t,q) + f_{LR}(t,q) + f_{UR}(t,q) + f_{RT}(t,q) + f_{Mention}(t,q). \tag{3.8}$$

Où $f_{MFR}(t,q)$ est le score global d'un tweet t en réponse à une requête q, f_{FR}, f_{LR}, f_{UR}, f_{RT} et $f_{Mention}$ sont les scores de pertinence partiels du tweet t suivant les critères que nous avons présentés.

Dans une seconde étape les auteurs ont pondéré l'ensemble des facteurs de pertinence pour leur associer des poids d'importance différents. Ainsi le score global $f_{WMFR}(t,q)$ de pertinence est obtenu comme suit :

$$f_{WMFR}(t,q) = w_0 * f_{FR}(t,q) + w_1 * f_{LR}(t,q) + w_2 * f_{UR}(t,q) + w_3 * f_{RT}(t,q) + w_4 * f_{Mention}(t,q) \tag{3.9}$$

Où $\sum_{i=1}^{4} w_i = 1$, cas dans lequel les auteurs ont montré que les résultats sont meilleurs. Le modèle proposé a été évalué en utilisant la mesure précision, dans une collection de tweets manuellement collectée.

Berardi *et al.* (2011b) ont proposé un système appelé *CipCipPy* pour la recherche de tweets qui exploite des mesures de qualité de texte pour filtrer le vocabulaire utilisé dans les tweets, et d'autres qui utilisent les informations contenues dans les *hashtags*. Les scores de chaque critère ont été combinés avec une simple combinaison linéaire où à chaque facteur est affecté un

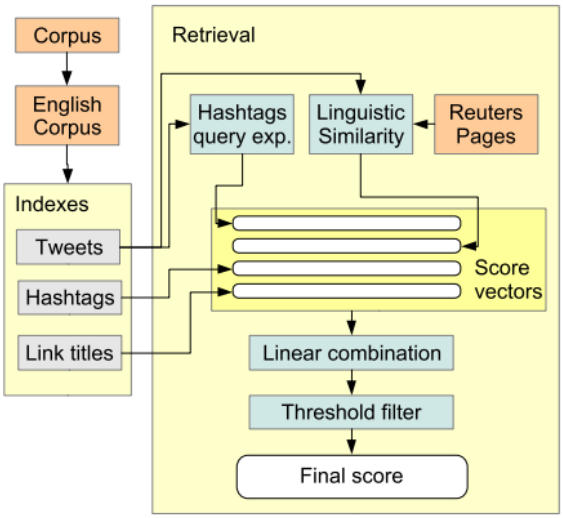


FIGURE 3.7: Architecture générale du système *CipCipPy* pour la recherche de tweets.

poids d'importance relatif. L'architecture générale du modèle proposé et le mécanisme de combinaison utilisé sont illustrés par la figure 3.7.

Les auteurs se sont aussi basés sur une variante du modèle BM25 pour le calcul du score thématique, donnée par la formule suivante :

$$score(d,q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{(1/TF(q_i,d)) \cdot (k_1+1)}{(1/TF(q_i,d)) + k_1 \cdot (1-b+b \cdot \frac{|d|}{avgdl})} \tag{3.10}$$

Où $avgdl$ est la longueur moyenne des documents de la collection, et k_1 et b sont les paramètres de la fonction BM25 déjà définis dans le Chapitre 2. Les auteurs ont trouvé que la mesure BM25 n'est pas adaptée pour la tâche Microblog, à cause la distribution compacte de la longueur des textes.

Dans la même direction de recherche, Damak *et al.* (2011, 2013) ont proposé un modèle de recherche intégrant divers dimensions de pertinence. Le premier exploite des facteurs basés sur le contenu (popularité et longueur des tweets), des facteurs liés à Twitter (fréquence/présence d'*URL*, hashtags) et d'autres liés aux auteurs des tweets (nombre de mentions, de tweets, etc). Pour le calcul du score final des tweets, les auteurs ont adopté une straté-

gie de combinaison linéaire des critères avec le score thématique obtenu par Lucene [1] :

$$score(d,q) = \alpha * lucene(d,q) + (1-\alpha) f_{critères}(d,q) \tag{3.11}$$

Où $f_{critères}(d,q)$ est le score global obtenu en combinant simplement et sans pondérations les différents critères considérés. Plus tard, les auteurs ont effectué une analyse des différents facteurs exploités et ont montré que l'emploi des URLs et l'expansion de requêtes a partir du *feedback* sont primordiales pour la RI dans les microblogs. Ils ont aussi constaté que le modèle probabiliste est plus performant que le modèle vectoriel en termes de précision (i.e., retourne plus de tweets pertinents).

Dans un contexte de recherche de tweets, Jabeur *et al.* (2012) ont aussi proposé des facteurs de pertinence qui incluent l'importance des auteurs et la fraîcheur des tweets. L'importance sociale est considérée comme un indicateur de crédibilité et réfère à l'influence des microbloggeurs dans le réseau social. La dimension temporelle est estimée en se basant sur les tweets qui sont temporellement proches pour présenter des termes similaires. Ainsi, les auteurs ont proposé un modèle bayésien dans lequel ils ont associé à chaque tweet t_j trois variables aléatoires t_{kj}, t_{oj} et t_{sj} selon trois contextes différents. La première variable t_{kj} correspond à la probabilité d'observer le tweet selon l'évidence thématique. La seconde variable t_{oj} correspond à la probabilité d'observer le tweet avec la connaissance implicite du contexte temporel. Enfin, t_{sj} correspond à la probabilité d'observer le tweet avec la connaissance implicite du contexte social. Ainsi, ces probabilités permettent de décomposer l'événement et d'observer un document selon les trois évidences : thématique, temporelle et sociale. Le score global de pertinence est donné par :

$$P(t_j|q) \propto \sum_{\forall \vec{k}} P(t_j|\vec{k}) P(t_j|\vec{k}) P(\vec{k}) \tag{3.12}$$

Où $\vec{k}$ est une configuration des termes inclus dans l'index, $P(t_j|\vec{k})$ est la probabilité qui dépend des trois sources d'évidences présentées. Cette probabilité est estimée comme suit :

$$P(t_j|\vec{k}) = P(t_{kj}|\vec{k}) P(t_{oj}|\vec{k}) P(t_{sj}|\vec{k}) \tag{3.13}$$

1. `http://lucene.apache.org`

Dans un contexte de recherche d'opinions, Gerani *et al.* (2012) ont analysé le problème d'incompatibilité des scores dans les scénarios d'agrégation multicritères et ont montré la nécessité de la transformation des scores avant d'appliquer la combinaison linéaire. La méthode proposée permet de générer des scores globaux de pertinence qui ne requièrent pas le fait que les scores individuels à combiner ne soient pas forcément comparables. Les auteurs se sont basés sur l'algorithme d'Espérance Conditionnelle Alternée (ACE) et l'algorithme BoxCox pour analyser le problème d'incomparabilité et effectuer une transformation de scores quand il est nécessaire. Une fois les fonctions g_t et g_o de transformation apprises en utilisant les scores de pertinence thématique et d'opinion (o), le score global est donné par :

$$score(q,d) = \alpha * g_t(p(q|d)) + (1-\alpha) * g_o(p(o|q,d)) \tag{3.14}$$

Où α est un paramètre dans $[0,1]$, qui a été obtenu empiriquement sur un ensemble d'apprentissage.

En RI personnalisée, la plupart des méthodes se basent sur des modèles de combinaison des scores originaux (algorithmiques) et des scores personnalisés des documents, calculés en fonction de leurs degré de similarité avec les profils utilisateurs représentant leurs centres d'intérêts (Gauch *et al.*, 2003; Sieg *et al.*, 2007). (Gauch *et al.*, 2003; Sieg *et al.*, 2007; Daoud *et al.*, 2010) ont proposé des modèles de combinaison des scores originaux des documents et des profils utilisateurs se basant sur des approches linéaires ou de produits de facteurs affectant des poids d'importance différents pour les deux critères. Parmi les approches utilisant un produit de facteurs pour l'agrégation, le travail de (White *et al.*, 2005), dans lequel les auteurs exploitent le profil utilisateur dans la chaîne d'accès à l'information. Pour intégrer les centres d'intérêts dans la fonction d'appariement du modèle de RI et calculer un score global de pertinence, les produit de facteurs est appliqué entre le score de correspondance des centres d'intérêts et le score de correspondance thématique entre la requête utilisateur et les documents. Dans le domaine de RI mobile, Cantera *et al.* (2008) proposent d'utiliser le modèle MCI (Multiplicative Competitive Interaction) comme modèle de combinaison des scores de la correspondance thématique d'un document, de la localisation géographique de l'utilisateur et de ses centres d'intérêts et ce, dans un cadre de RI mobile. Ainsi, l'expression générale d'utilité d'un document selon le modèle MCI est donnée par une combinaison linéaire des scores individuels. Les attributs contextuels considérés sont principalement la localisation et le contexte du dispositif mobile qui sont combinés avec le score textuel des documents.

Tâche de recherche	**Travaux**	**Critères de pertinence**	**Opérateurs d'agrégation**
RI mobile	(Church et Smyth, 2008; Hattori *et al.*, 2007; Cheverst *et al.*, 2000; Schilit *et al.*, 2003; Yau *et al.*, 2003; Cantera *et al.*, 2008)	Thématique, centres d'intérêts, localisation géographique, temps, descripteurs sociaux	Mécanismes de combinaison linéaire
RI personnalisée	(Liu *et al.*, 2004; Ma *et al.*, 2007; Sieg *et al.*, 2007)	Aboutness, coverage, appropriateness, reliability, centres d'intérêts	Mécanismes de combinaison linéaire : somme des scores partiels, produit de facteurs
RI sociale	(Becker *et al.*, 2011; Metzler et Cai, 2011; Berardi *et al.*, 2011b; Chen *et al.*, 2012; Smith *et al.*, 2008; Leung *et al.*, 2006)	contenu, descripteurs de pertinence liée à Twitter, autorité, temps	Mécanismes de combinaison linéaire : somme des scores partiels, produit de facteurs
RI géographique	(Mata et Claramunt, 2011; Kishida, 2010; Daoud et Huang, 2013)	Contenu, temps, localisation géographique, proximité	Mécanismes de combinaison linéaire : somme des scores partiels

TABLE 3.2: Synthèse des travaux impliquant l'agrégation de pertinence multidimensionnelle.

Dans la même direction de recherche, Cong *et al.* (2009) ont proposé un modèle de RI basé sur la localisation géographique et le critère thématique

dans lequel les documents sont classés par le biais d'une combinaison linéaire des deux dimensions. Le tableau 3.2 présente une synthèse des travaux qui se sont intéressés au problème de combinaison de la pertinence multicritères suivant les cadre de RI mentionnés ci-dessus.

3.5.1.2 Moyennes ordonnées

Malgré son exploitation massive dans divers domaines de recherche tels qu'en intelligence artificielle, dans le domaine mathématiques appliquées et surtout en logique floue, l'opérateur OWA (Yager, 1988) a été très peu utilisé en RI. Boughanem *et al.* (2006) ont proposé une méthode basé sur le même principe que OWA. Les auteurs ont exploité l'opérateur OWmin, qui utilise un vecteur de pondération pour minimiser l'impact des termes ayant les plus faibles degrés sur la valeur finale. Ainsi, de même que pour les OWA, les vecteurs de degrés sont triés et pondérés, mais c'est le minimum des valeurs qui est ensuite considéré, et non leur moyenne. Deux méthodes de pondérations ont été considérées, l'une basée sur l'implication de Dienes, l'autre sur l'implication de Gödel . Pour un vecteur $T = t_1, \ldots, t_n$ représentant les degrés résultats pour un document, t_i est le degré de possibilité entre le terme i de la requête et le document, une fois le vecteur trié de manière décroissante. Soit le vecteur de pondération $W = w_1, \ldots, w_n$. L'agrégation par OWmin utilisant l'implication de Dienes est donnée par :

$$OWmin(T, W) = min_i(max(t_i, 1 - w_i)) \tag{3.15}$$

Dans cette formulation, le poids est considéré comme un niveau d'importance pour les degrés. L'utilisation de l'implication de Gödel donne :

$$OWmin(T, W) = min_i(w_i \rightarrow t_i) \tag{3.16}$$

Où l'implication est définie par $w_i \rightarrow t_i = 1$ si $w_i \leq t_i$, t_i sinon.

L'approche proposée est évaluée sur un ensemble de la collection CLEF2001, et les résultats expérimentaux ont montré l'efficacité de la méthode comparativement aux méthodes classiques se basant sur les opérateurs *min*.

3.5.1.3 Approches de surclassement

Parmi les premiers modèles multicritères de surclassement proposés en RI, on cite celui de Farah et Vanderpooten (2006). Dans ce modèle, Farah et Vanderpooten (2006) ont développé les critères de pertinence en se basant sur les différents facteurs (g_i) qui ont un impact sur la définition de la pertinence des documents. Parmi ces critères, nous citons :

- Fréquence du terme : pour les requêtes contenant un seul terme (i.e., $q = t_k$), $g_1(d, t_k) = \frac{tf_k}{max_{tf}}$, pour les requêtes contenant plusieurs termes, l'opérateur moyenne est utilisé pour agréger les valeurs.
- Position du terme : pour les requêtes contenant un seul terme, $g_2(d, t_k) = \sum_{a=1}^{4} l_{k,a}$, où $l_{k,a}$ est une valeur binaire qui est égale à 1 si le terme t_k apparaît dans l'emplacement l_a du document d, 0 sinon. Ces emplacements sont l'URL (l_1), le titre (l_2), les mots clés (l_3) et la description (l_4).
- Autorité : le nombre de liens entrants au document d.
- proéminence : le nombre de documents "fils" de d (i.e., documents qui apparaissent dans un niveau hiérarchique inférieur à d selon le *sitemap* de d
- Proximité : proximité des termes de la requête dans le document
- Longueur du document
- Rareté : le nombre de documents dans lesquels apparaissent les termes de la requête.

Dans ce modèle multicritères, l'ordonnancement est obtenu en deux étapes. Dans la première étape, les auteurs ont appliqué des règles de décision simples dans des comparaisons par paires des documents. Ceci permet de juger si un document est globalement plus ou moins pertinent qu'un autre selon leur performances sur les critères retenus. Plusieurs niveaux d'exigence peuvent être définis sur les règles de décision, ce qui permet la construction de plusieurs relations binaires entre les paires de documents. Par exemple, une règle de décision simple consiste à accepter qu'un document d est globalement "au moins aussi bon que" d' lorsque $g_j(d) \geq g_j(d') - s_j$, $\forall j$, où s_j est un seuil d'indifférence permettant de modéliser l'imprécision de l'évaluation de la performance d'un document selon le critère g_j. Ces relations binaires sont ensuite exploitées pour obtenir l'ordonnancement final.

Farah et Vanderpooten (2006) ont évalué le modèle de surclassement en utilisant les requêtes TD ("Topic Distillation") de la tâche Web de TREC 2004 (Craswell et Hawking, 2004). Le modèle a montré des bonnes performances en comparaison avec des approches d'agrégation classiques telles que

les opérateurs *somme*, *produit*, *max* et *min*.

3.5.1.4 Approches d'agrégation prioritaires

Les travaux de (da Costa Pereira *et al.*, 2009, 2012) étaient parmi les premiers travaux sur l'agrégation multicritères en RI et sur les opérateurs d'agrégation prioritaires en particulier. Les auteurs ont proposé une approche d'agrégation multidimensionnelle mettant en jeu quatre critères des pertinence : contenu, couverture, adéquation et fiabilité en définissant deux opérateurs d'agrégation prioritaire en l'occurrence, *Scoring* et *And*. Ces opérateurs modélisent un ordre de priorité entre les critères de pertinence sur la base du mode de calcul de poids associés qui favorise la satisfaction du critère d'ordre supérieur.

L'opérateur "Scoring". Dans cet opérateur, le poids de chaque critère dans le score global dépend à la fois des poids et des scores de satisfaction des critères les plus prioritaires : plus le score de satisfaction des critères de plus haute priorité est élevé, plus le score de satisfaction d'un critère de moindre priorité est moins influençable dans le score global d'un document. L'instanciation de cet opérateur sur l'approche d'agrégation prioritaire, que nous avons déjà présenté dans la section 3.3.3, est appliquée de la façon suivante :

- Les alternatives sont considérés comme un ensemble de documents $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$
- L'ensemble des critères $\mathcal{C}$ représentent les dimensions de pertinence
- Les évaluations $v_{d_1}, \ldots, v_{d_n}$ représentent les scores partiels des documents par rapports aux critères, où chaque v_{d_i} représente le *rsv* du document d par rapport au critère c_i.

La fonction d'agrégation $g(v_1(d), \ldots, v_n(d))$ représente le score global du document par rapport à tous les critères. $g : [0,1]^n \rightarrow [0,n]$ est calculé comme suit :

$$g(v_1(d), \ldots, v_n(d)) = \sum_{i=1}^{n} \lambda_i \cdot v_i(d) \tag{3.17}$$

L'opérateur "And". Cet opérateur est inspiré de l'opérateur d'agrégation classique "*min*". L'idée de cet opérateur consiste à défavoriser les critères les moins satisfaits s'ils ne sont pas pas importants ou prioritaires. Par exemple, si on considère un document d, pour lequel le critère c_i ayant le score de satisfaction le plus bas, est au même temps le critère le moins important, alors le score global de satisfaction du document sera plus grand que le

score partiel $v_i(d)$. Contrairement à l'opérateur "*min*", où c'est $v_i(d)$ qui aurait été considéré. Le score global de pertinence donné par l'opérateur d'agrégation "And" est défini par :

$$g(v_1(d), \ldots, v_n(d)) = min_{i=1,n}(\{v_i(d)\}^{\lambda_i}) \tag{3.18}$$

De cette façon, moins un critère est important, moins il a de chance pour représenter le score global de satisfaction du document.

Les premiers travaux basés sur les deux opérateurs "And" et "Scoring" sont ceux de (da Costa Pereira *et al.*, 2009, 2012) qui ont proposé une représentation multidimensionnelle de la pertinence et ont exploité 4 critères dans un cadre de RI personnalisée à savoir, *aboutness, coverage, appropriateness* et *reliability*. Les critères *coverage* et *appropriateness* sont explicitement liés à la représentation formelle du profil utilisateur qui est représenté comme un sac de mots. De plus, la personnalisation est liée à la tâche de recherche, i.e., si un utilisateur préfère un critère à un autre une priorité sur les deux critères doit être spécifiée. De cette manière, l'évaluation d'une même requête pour le même utilisateur (ou pour des utilisateurs différents) peut produire des ordonnancements de documents différents. Un des avantages de ces approches, est qu'ils permettent aussi l'identification des poids d'importance des critères sans aucun processus d'apprentissage. Bouidghaghen *et al.* (2011b) ont appliqué ces deux opérateurs dans un modèle multicritères de RI mobile basé sur trois dimensions à savoir, la thématique de recherche, les centres d'intérêts des utilisateurs et la localisation géographique. La méthode proposée a été évaluée dans un cadre d'évaluation intégrant l'utilisateur et son contexte (30 profils utilisateurs) dans une démarche d'évaluation basée-simulation. Les auteurs ont défini six scénarios d'évaluation selon l'ordre de priorité que peut choisir un utilisateur sur les trois critères de pertinence. L'évaluation expérimentale a montré que la considération des trois dimensions de pertinence contextuelle améliore la pertinence des résultats par rapport à la dimension thématique seule. Ainsi, La méthode de combinaison linéaire pondérée a atteint les meilleures performances en termes de $nDCG$ et de précision en comparaison avec les méthodes d'agrégation classiques. Les auteurs ont aussi conclu que l'opérateur "Scoring" améliore l'opérateur "And" et la moyenne pondérée (*weighted average*).

3.5.2 Approches basées sur l'agrégation de listes

3.5.2.1 Approches d'agrégation d'ordonnancements

Une panoplie de modèles d'agrégation d'ordonnancements a été proposée dans la littérature en RI. La plupart des approches exploitent en particulier les descripteurs thématiques des documents tels que la longueur, la fréquence des termes ou des modèles de recherches complètement indépendants (Farah et Vanderpooten, 2007, 2008).

Lee (1997) a montré que les méthodes *CombSUM* et *CombMNZ* sont plus performantes que les autres méthodes de combinaison *Comb∗*. L'étude a été effectuée sur les résultats de 6 participants à la tâche ad-hoc de TREC3. Récemment, une autre étude menée par (Cormack *et al.*, 2009) a montré que la méthode de combinaison *Reciprocal Rank Fusion* (RRF) est meilleure que *CombMNZ* et Condorcet. Cormack *et al.* (2009) ont comparé les approches en combinant les résultats de 30 systèmes participants à TREC sur les topics 351 – 400 ainsi que sur les résultats des modèles de la tâche ad-hoc de TREC 3, TREC 5 et TREC 9. Les auteurs ont aussi évalué RRF sur la collection LETOR [2] 3.

Aslam et Montague (2001) ont proposé la méthode *Borda Fuse*, qui peut être perçue comme une variante pondérée de la méthode *Borda Count* pour la méta-recherche. Plus particulièrement, l'approche associe des poids différents pour les modèles de recherche, où les poids sont calculés en se basant sur les moyennes des mesures de précision données par les modèles d'ordonnancements. L'évaluation de *Borda Fuse* sur les résultats des systèmes participant à la tâche ad-hoc de TREC 3, TREC 5 et TREC 9 montre que cette dernière est plus performante que *Borda Count*. En effet, l'un des inconvénients majeurs des modèles basés sur *Borda Count* est qu'ils attribuent une importance égale à tous les systèmes de recherche, ce qui n'est pas toujours réaliste.

Greengrass (2000) a proposé le modèle *Round Robin* permettant d'agréger des documents selon leurs rangs dans les listes de résultats. Le premier document dans la liste finale sera le premier document dans la première liste retournée et ainsi de suite. Le choix des listes se fait de manière aléatoire. Si et Callan (2003) affirment que *Round Robin* est un choix à opérer si les scores individuels des documents sont complètement incomparables (lorsque

2. `http://research.microsoft.com/en-us/um/beijing/projects/letor/letor3dataset.aspx`

les systèmes de recherche utilisent différents algorithmes ou lorsque les collections utilisent différentes statistiques). Cependant, les travaux de (Renda et Straccia, 2003) ont montré que les meilleurs résultats de *Round Robin* sont obtenus lorsqu'il s'agit des mêmes stratégies de recherche. Cependant, comme pour les méthodes se basant sur *Borda Count*, *Round Robin* attribue une importance égale à tous les systèmes de recherche, même si un système peut être moins pertinent que d'autres relativement à cette requête.

Plus tard, dans le contexte de méta-recherche, (Ah-Pine, 2008) s'est penché sur le problème d'agrégation d'ordonnancements en utilisant des opérateurs d'agrégation non linéaires définis dans le domaine de fusion de données, tels que les opérateurs *T-norms* (Menger, 1942) et *T-conorms*. Les expérimentations menées sur une base de test réelle montrent que ces opérateurs donnent des meilleurs résultats que les méthodes *combSUM* et *combMNZ*.

3.5.2.2 Approches d'apprentissage d'ordonnancements

En RI, ces approches ont été largement exploitées dans le cas où l'on dispose d'un grand nombre de descripteurs de pertinence. Ces descripteurs se réfèrent généralement à la dimension de pertinence thématique. Le tableau 3.3 présente les descripteurs de pertinence les plus souvent utilisés avec les algorithmes d'apprentissage d'ordonnancements.

Dans les tâches de recherche de tweets, ces algorithmes ont montré des bonnes performances pour la combinaison des critères de pertinence sur Twitter (Duan *et al.*, 2010; Metzler et Cai, 2011; Chang *et al.*, 2013). Duan *et al.* (2010) ont identifié des critères basés sur le contenu comme la longueur des tweets et la présence ou non des *URL*s et d'autres qui reflètent l'importance de l'utilisateur tel que l'autorité, ont analysé l'effet de chaque facteur à part dans l'évaluation globale de pertinence. La liste des descripteurs de pertinence exploitée comprend :

- *URL* : si le tweet contient des URL
- Nombres d'*URL* dans la collection
- Nombre de *retweet*
- Score des *hashtags* : somme des fréquences des top-*n* *hashtags* qui apparaissent dans le tweet
- Réponse : si le tweet en question est une réponse à un autre tweet
- *OOV (Out Of Vocabulary)* : ratio des mots ne faisant pas partie du vocabulaire.

Descripteur	**Explication**	**Niveau**
Nombre d'occurrence	Nombre de fois une requête apparaît dans le titre, les ancres, l'URL, le titre extrait, corps du texte	Modèle
BM25	Les scores BM25 sur le titre, les ancres, l'URL, le titre extrait, requête, et corps du texte	Modèle
N-gram BM25	Les scores N-Gram BM25 sur le titre, les ancres, l'URL, le titre extrait, requête, et corps du texte	Modèle
Distance de similarité	Score de la distance de similarité entre la requête et le titre, les ancres, l'URL, le titre extrait, requête, et le corps du texte	Modèle
Liens entrants	Nombre de liens entrants à la page	Document
PageRank	L'importance de la page calculé dans un graphe Web	Document
Nombre de cliques	Nombre de cliques sur une page dans un log de recherche	Document
BrowseRank	Le score d'importance d'une page calculé dans un graphe de navigation d'utilisateurs	Document
Score de qualité d'une page	Vraisemblance d'une page de mauvaise qualité	Document
Score de spam	Vraisemblance d'une page spam	Document

TABLE 3.3: Exemples de descripteurs de pertinence utilisés par les méthodes d'apprentissage d'ordonnancements pour la RI (Li, 2011).

Les auteurs ont utilisé l'algorithme RankSVM pour l'apprentissage du modèle d'ordonnancement en se basant sur tous les descripteurs présentés. Une des observations qui est ressortie de cette étude est que le critère présence d'*URL* est important pour le modèle, alors que les critères *hashtags* et *retweets* ne sont pas discriminants.

Metzler et Cai (2011) a également proposé d'exploiter l'algorithme d'apprentissage par paire ListNet, se basant sur un critère de similarité textuel, un critère dépendant du temps pour mesurer la fraîcheur des documents et un ensemble de caractéristiques Twitter tels que la présence d'*URL* et de *hashtags*, etc. Étant donnés une requête q et un tweet t, le modèle calcule le

score de pertinence $score(q,t)$ suivant la formule suivante :

$$score(q,t) = \sum_{i=1}^{n} \lambda_i f_i(q,t) \tag{3.19}$$

Où n est le nombre de descripteurs de pertinence, $f_i(q,t)$ est la fonction calculant la pertinence suivant le descripteur i et λ_i est un paramètre du modèle. L'algorithme a donné des bons résultats dans le cadre de la tâche Microblog de TREC. Plusieurs autres efforts ont été déployés dans l'application des algorithmes d'apprentissage d'ordonnancements dans la recherche de tweets. Miyanishi *et al.* (2012) ont appliqué un algorithme non spécifié en classant les tweets pour chaque topic. Cheng *et al.* (2013) ont proposé une revue critique des méthodes d'apprentissage d'ordonnancements dans le cadre de recherche temps réel sur Twitter. Les auteurs ont évalué empiriquement de nombreuses techniques de l'état de l'art en utilisant divers descripteurs de pertinence :

– Critère basé sur le contenu
– Richesse du contenu
– Fraîcheur : relation temporelle entre le temps de soumission des requêtes et le temps de publication des tweets
– Autorité : influence des auteurs
– Critères liés aux tweets (*retweets*, mentions, *hashtags*)

L'évaluation expérimentale des algorithmes RankSVM et LambdaMART sur la collection de tweets fournie par la tâche Microblog de TREC 2011 ont montré que l'algorithme RankSVM donne des meilleurs résultats dans la plupart des cas.

Le tableau 3.4 catégorise les méthodes d'apprentissage d'ordonnancements, indépendamment du contexte dans lesquels ils ont été appliquées. Nous présentons les algorithmes les plus importants qui sont basés sur les SVM, la technique de *boosting* et les réseaux de neurones, respectivement.

La différence majeure entre toutes ces méthodes réside en grande partie dans la manière avec laquelle la fonction de perte est définie. Il est aussi important de noter que les méthodes par listes sont les plus adaptées pour les problèmes liés à l'ordonnancement en RI (Liu, 2009; Li, 2011).

Le tableau 3.5 présente une petite synthèse sur les principaux avantages et inconvénients de ces méthodes. Les algorithmes d'apprentissage d'ordonnancements ont montré des bonnes performance dans plusieurs tâches de

	SVM	**Boosting**	**Réseaux de neurones**
Point	OC SVM	McRank	-
Paire	Ranking SVM, IR SVM	RankBoost, GBRank, LambdaMART	RankNet, Frank, LambdaRank
Liste	SVM MAP, PermuRank	AdaRank	ListNet, ListMLE

TABLE 3.4: Catégorisation des méthodes d'apprentissage d'ordonnancements.

Approche	**Modèle**	**Avantages**	**Inconvénients**
Point	Régression, Classification, Régression ordinale	Facilité d'exploiter et optimiser les théories et algorithmes existants	Ordonnancement précis, Score ou catégorie précis, information sur la position est invisible à la fonction de perte
Paire	Classification par paire		
Liste	Ordonnancement	Basé sur les requêtes et les positions	Plus complexe, nouvelles théories nécessaires

TABLE 3.5: Avantages et inconvénients des méthodes d'apprentissage d'ordonnancements.

recherche Liu (2009); Li (2011); Macdonald *et al.* (2013). Cependant, malgré leur popularité, ces méthodes présentent plusieurs inconvénients. Par exemple, la plupart d'entre elles ne tiennent pas compte des dépendances inter-documents et entre les descripteurs de pertinence (Chapelle *et al.*, 2011). De plus, ces méthodes ont tendance à offrir un aperçu limité sur la façon de considérer l'importance et l'interaction entre les groupes de caractéristiques qui représentent les différentes dimensions de pertinence. Il est très difficile pour un décideur de comprendre pourquoi un critère est préféré par rapport à un autre (Eickhoff *et al.*, 2013a). Un autre problème est la complexité des algorithmes et le temps d'exécution pour apprendre ou pour tester un ensemble de requêtes. En effet, le grand nombre de ca-

ractéristiques utilisées par ces approches peut influencer négativement la qualité du classement ainsi que la vitesse d'exécution des algorithmes. Par ailleurs, beaucoup de caractéristiques sont utilisées par les algorithmes d'apprentissage d'ordonnancements, alors qu'elles peuvent être non pertinentes ou bruitées.

Ces problèmes ont été aussi soulevés dans "*The Yahoo ! Learning to Rank Challenge*" Chapelle *et al.* (2011), un défi dans lequel plusieurs groupes de recherche participent avec des algorithmes d'apprentissage d'ordonnancements évalués sur une base de test standard fournie par Yahoo [3]. Parmi les leçons tirées de ce défi, est que malgré la maturité et la complexité de ces algorithmes, il s'avère qu'ils sont tous moins performants qu'un simple référentiel de comparaison, donné par les organisateurs, qui est basé sur un modèle de régression. Les organisateurs de la tâche expliquent ce fait par Chapelle *et al.* (2011) :

> "*There are two possible explanations for this. One of them is that some of the "improvements" reported in papers are due to chance. A recent paper (Blanco et Zaragoza, 2011) analyzes this kind of random discoveries on small datasets. The other explanation has to do with the class of functions. In general, the choice of the loss function is all the more critical as the class of function is small, resulting in underfitting.*"

3.6 Conclusion

Dans ce chapitre, nous avons présenté le problème d'agrégation multicritères et nous l'avons situé dans le cadre de la RI. Nous avons montré les différentes approches proposées dans la littérature pour la combinaison de pertinence multidimensionnelle en RI. Nous avons formalisé les différentes techniques proposées et nous les avons classés selon la manière avec laquelle les scores (ou les préférences) sont calculés. Nous pouvons noter que de nombreux travaux ont exploité différents critères ou descripteurs de pertinence, mais très peu se sont intéressés au problème spécifique de la construction de la fonction d'agrégation de pertinence. Dans le chapitre suivant, nous nous intéressons aux approches d'agrégation de pertinence multicritères qui ont particulièrement exploité le critère temporel dans le processus d'agrégation

3. `http://learningtorankchallenge.yahoo.com/datasets.php`

et qui ont été proposées dans des collections dynamiques sous forme de flux de de données. Nous allons montrer également l'importance du critère temporel dans tous les niveaux d'un processus typique de RI.

Chapitre 4

Recherche d'information temporelle et pertinence : synthèse des travaux de l'art

---

"*Time is a ubiquitous factor at many stages in the information-seeking process, with users having temporally-relevant information needs, and collections having temporal properties at collection, document metadata, and document content levels*".
Derczynski *et al.* (2015)

4.1 Contexte et problématique

Dans le chapitre 3, nous avons montré que la pertinence est un concept multidimensionnel et nous avons présenté les approches multicritères permettant d'agréger les différents critères entrant en jeu dans le processus final d'ordonnancement des documents. Nous avons aussi souligné l'importance de la pertinence thématique par rapport aux autres critères qui dépendent en grande partie de l'environnement utilisateur (e.g., localisation géographique, dispositif physique) et du contenu des documents (e.g., *hashtags*, présence d'*URL* dans le cas des microblogs, etc). La dimension temporelle est éga-

lement très importante dans le jugement de pertinence des documents, elle demeure une dimension primordiale surtout dans les cadres de RI où les utilisateurs s'intéressent à des documents récents ou à des informations issues des flux de données. Cependant, la plupart des travaux de l'état de l'art exploitent ce critère, en plus d'autres facteurs, d'une façon brute en se basant sur le temps de soumission des requêtes utilisateurs ou sur le temps de publication des documents. De plus, les facteurs (ou descripteurs) de pertinence exploités et les mécanismes d'agrégation utilisés sont majoritairement appliqués sur des collections de documents statiques. Un modèle proposé dans ce cadre peut donner des bons résultats en moyenne, mais pourrait aussi être très mal adapté quand il s'agit d'une collection de documents qui change dans le temps (Wang *et al.*, 2003; Harper et Chen, 2012). En effet, de nombreuses problématiques sont soulevées par l'introduction du critère temporel dans le processus de recherche d'information :

1. Comment représenter le critère temporel et quels sont les niveaux dans lesquels le temps pourrait être injecté (documents, requêtes, modèle d'ordonnancement) ?
2. Comment identifier la sensitivité des requêtes au temps ?
3. Quel mécanisme de combinaison multicritère utiliser pour agréger les facteurs de pertinence dans une collection de documents qui évolue dans le temps ?

Pour répondre à ces enjeux, plusieurs approches, détaillées dans les sections suivantes, sont proposées dans la littérature. Nous les classons en trois catégories (Moulahi *et al.*, 2015c) :

– ***Les approches exploitant le temps au niveau des requêtes*** : l'objectif est d'éliciter l'aspect temporel des intentions de recherche des utilisateurs à partir des requêtes et paramétrer le modèle en se basant sur le type des requêtes identifiées. Un des objectifs phares à ce niveau consiste à identifier la période à laquelle une requête fait référence. Une approche sensible au temps doit aussi être en mesure de faire face aux requêtes ambiguës (i.e., pouvant faire référence à plusieurs périodes de temps). Parmi les applications qui ont été proposée à cette fin celles abordant la dynamicité des requêtes (Vlachos *et al.*, 2004; Jones et Diaz, 2007; Metzler *et al.*, 2009; Subasic et Castillo, 2010; Kulkarni *et al.*, 2011; Osborne *et al.*, 2012; Radinsky *et al.*, 2012). Par exemple, pour certaines requêtes, Google adapte ses résultats de recherche en fonction du type de la requête et de la période du temps dans laquelle elle a été soumise. Considérons la requête

"Karim Benzema", soumise le 29 Juin 2014, pendant la coupe du monde de football de 2014, Google affiche en premier lieu, comme montré dans la figure 4.1, des statistiques concernant le joueur pendant la compétition. Ensuite, il affiche quelques actualités puis la page Wikipedia concernant "Benzema".

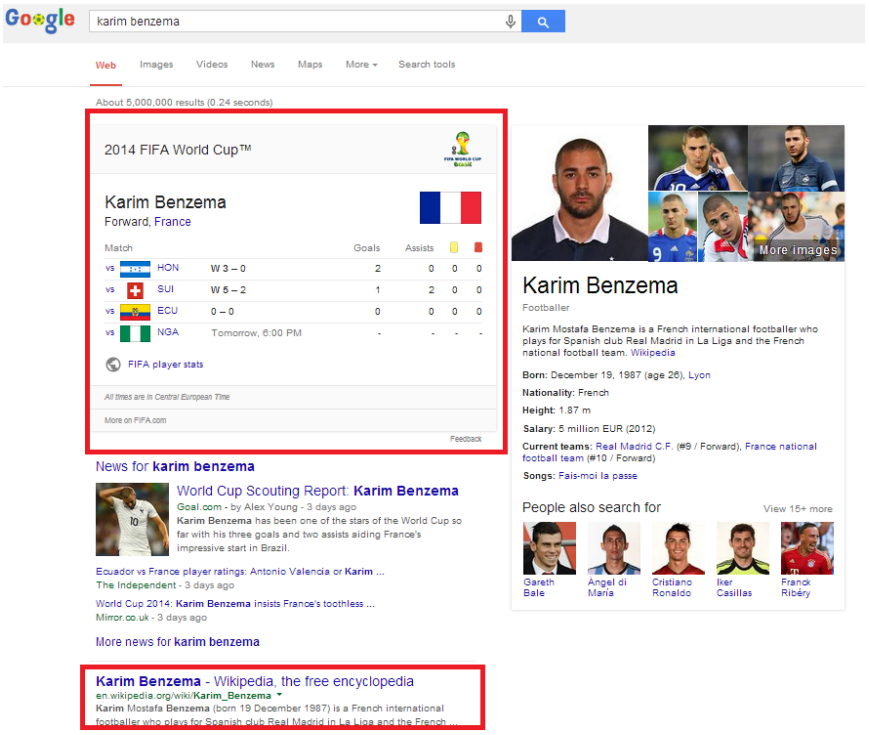


FIGURE 4.1: Résultats de recherche sur Google pour la requête "Karim Benzema", soumise le 29/06/2014.

- ***Les approches exploitant le temps au niveau des documents*** : visent à identifier et extraire, dans un premier temps, les expressions temporelles contenues dans les documents. Cette étape permet de déterminer les dates aux quelles font référence un document (Alonso *et al.*, 2011; Manica *et al.*, 2012; Strötgen *et al.*, 2012). Ceci permet d'améliorer la tâche de recherche d'information dans une étape ultérieure.
- ***Les approches exploitant le temps au niveau des modèles d'ordonnancement*** : combinent la dimension temporelle avec d'autres dimensions de pertinence susceptibles d'améliorer les performances des SRI dans un cadre de recherche particulier. Dans cette catégorie, nous nous

intéressons aux approches qui ont été proposées dans des collections de documents qui changent dans le temps (Kanhabua et Nørvåg, 2010; Dakka *et al.*, 2012; Harper et Chen, 2012; Efron *et al.*, 2014; Lin *et al.*, 2014b).

Dans ce chapitre, nous commençons dans la section 4.2 par la définition de la notion du temps dans le cadre du RI. Nous donnons aussi une définition générale des concepts des critères récence et fraîcheur de l'information. Dans la section 4.3, nous proposons un schéma général pour catégoriser les travaux de l'état de l'art suivant la manière avec laquelle le temps est exploité. Ensuite, nous présentons ces travaux en fonction du cadre de RI temporelle associé. Nous présentons principalement les approches qui ont été proposées dans le cadre de la RI traditionnelle. La section 6.5 conclut le chapitre.

4.2 Notions préliminaires

Le temps peut être défini comme : "*une mesure avec laquelle on peut ordonner des événements du passé au présent et jusqu'au futur, et aussi une mesure des durées des événements et de l'intervalle entre eux*" [1]. Les événements peuvent être définis comme des faits ou des changements ordonnés dans le temps et présentés sous forme de textes, de tableaux, des graphiques et des journaux (*timelines*). En RI classique, le temps est généralement représenté par la date de création des documents, le temps de soumission d'une requête ou par les expressions temporelles qui sont contenues dans les documents. La dimension temporelle peut être exploitée suivant deux critères de pertinence :

- *La récence (recency).* C'est un critère qui permet de favoriser les documents récemment publiés (Dong *et al.*, 2010). Elle peut aussi faire référence à des requêtes pour lesquelles l'utilisateur s'attend à des documents qui sont à la fois pertinents et récents. Ce critère est souvent calculé comme la différence entre le temps de publication du document et le temps de soumission de la requête.
- *La fraîcheur d'information (freshness).* Peut être interprétée de différentes manières, en fonction du type des requêtes. Par exemple, pour les requêtes liées aux sujets d'actualités (*news*), la fraîcheur concerne principalement les documents qui traitent des *nouvelles* informations (Amati *et al.*, 2012). Toutefois, il convient de noter que la plupart des travaux

1. `http://en.wikipedia.org/wiki/Time`

en RI temporelle ne font pas une distinction claire entre la récence et la fraîcheur d'information.

Mathews et Kanmani (2012) ont identifié deux axes de recherche majeurs qui incluent les dimensions temporelles en RI : le domaine de recherche d'information sensible au temps (*Temporal Information Retrieval*) et l'ordonnancement temporel (*Temporal Ranking*). Les deux domaines ont récemment émergé, et visent à exploiter le temps pour extraire des documents temporellement pertinents. Dans le domaine de RI sensible au temps, les critères temporels les plus utilisés sont la date de publication et la pertinence thématique des documents. Dans le domaine de l'ordonnancement temporel, le temps est intégré directement dans le modèle de RI ou pris en considération à travers des modèles qui visent à éliciter le type des requêtes et ordonne les documents suivant ce type. Dans cette thèse, nous ne faisons pas de distinction entre les deux domaines.

4.3 Classification générale des approches de RI sensibles au temps

4.3.1 Aperçu général

La RI temporelle est un nouveau domaine de recherche dont l'objectif principal est l'amélioration des modèles de RI classiques en exploitant les informations temporelles qui peuvent exister dans les requêtes et les documents. D'une façon générale, les modèles de RI combinent la notion de pertinence traditionnelle avec la dimension temporelle. Toutefois, pour bien cerner le besoin des utilisateurs, certains travaux s'intéressent à l'étude de la temporalité des requêtes, pour identifier éventuellement les périodes auxquelles fait référence une requête. D'autres travaux tentent d'exploiter le temps au niveau des documents. Le premier défi ici consiste à identifier et extraire les expressions temporelles qui peuvent être contenues dans ces documents. Le deuxième défi est la représentation et la normalisation de ces expressions (Nunes *et al.*, 2008). Alonso *et al.* (2011) distingue trois types d'expressions temporelles : expressions explicites (e.g., 2015), implicites (e.g., Vacances de Noël) ou relatives (e.g., hier, vendredi, etc). La difficulté principale ici est de traiter les expressions temporelles relatives qui requièrent leur normalisation vers la date de création du document. Cependant, l'information concernant cette date n'est pas toujours disponible.

L'objectif final de toutes ces approches, exploitant des descripteurs temporels au niveau des documents et des requêtes est d'améliorer les performances des modèles de RI et d'adapter le besoin à caractère temporel des utilisateurs. Pour répondre à tous ces défis, les systèmes de RI sensibles au temps intègrent le critère temporel dans le modèle d'ordonnancement en le combinant avec d'autres critères de pertinence. La plupart des approches de littérature se basent sur des mécanismes de combinaison linéaire (critère thématique et temporel) ou des modèles de langue (Berberich *et al.*, 2010). Ces méthodes ont conduit à l'émergence de plusieurs applications de RI où le temps est au centre du modèle d'ordonnancement.

Pour donner un aperçu plus général des méthodes exploitant le temps, nous présentons dans la figure 4.2, une classification plus détaillée des approches s'intéressant au temps en RI suivant les trois niveaux précédemment discutés (Moulahi *et al.*, 2015c).

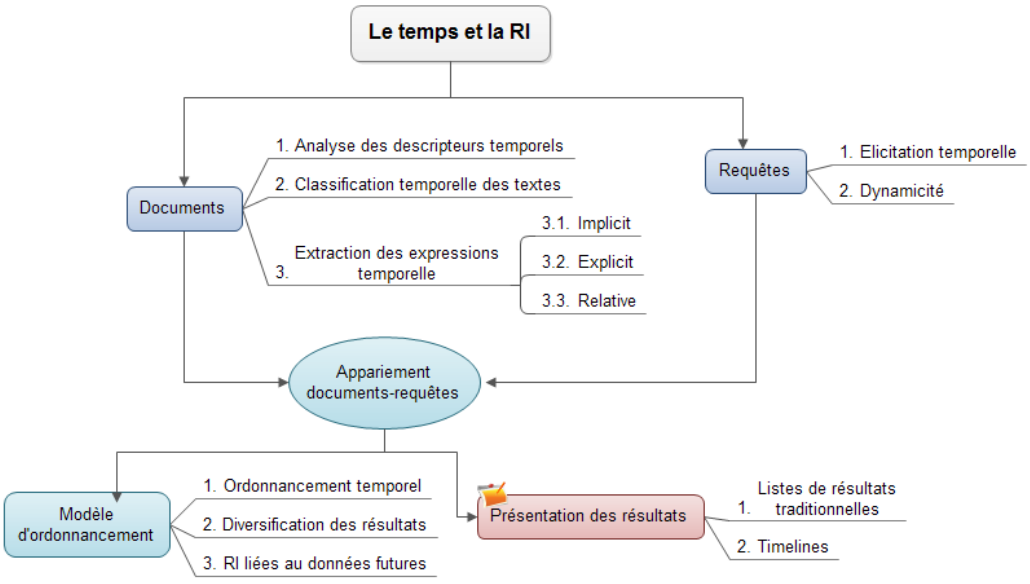


FIGURE 4.2: Classification des principaux axes de recherche s'intéressant au temps en RI suivant les trois niveaux : requête, document et modèle de RI (Moulahi *et al.*, 2015c).

Dans ce qui suit, nous présentons les travaux effectués en RI temporelle suivant le niveau dans lequel ils ont été appliqués. Plus spécifiquement, nous introduisons les approches ou le temps est exploité au niveau des requêtes,

documents et modèle d'ordonnancement.

4.3.2 Le temps au niveau de la requête

Les recherches récentes ont montré que plusieurs requêtes Web contiennent des expressions temporelles explicites ou implicites. Une analyse d'une archive des requêtes Web a montré que 7% des requêtes contiennent des intentions de recherche temporelles implicites (Metzler *et al.*, 2009). Jones et Diaz (2007) ont identifié trois types de profils temporels de requêtes : requêtes atemporelles, requêtes temporellement non ambiguës et requêtes temporellement ambiguës. Tandis que les requêtes atemporelles font référence à des des sujets non sensibles au temps (e.g., "Amazon"), les requêtes temporellement non ambiguës sont celles faisant référence à une période de temps précise (par exemple, "2015"). Les requêtes temporellement ambiguës se réfèrent à des périodes de temps différentes (e.g., "Ligue des champions"). Nunes (2007) a montré que seulement $1,5\%$ des requêtes sont soumises avec des expressions temporelles explicites. Il est donc très difficile de désambiguïser l'intention temporelle des ces requêtes. Par exemple, si on considère la requête "Michael Jackson", cette dernière peut être sur la biographie du chanteur ou des vidéos du chanteur, mais peut aussi concerner des nouvelles ou des articles le concernant qui sont liés à des événement récents (e.g., sa nouvelle chanson réalisée en 2014, anniversaire, etc). Si on reprend la requête "Karim Benzema" (Cf., Figure 4.1), dans une période où il n'y a pas d'événement concernant le joueur, Google affiche en première position, comme le montre la figure 4.3 les pages Wikipedia de "Benzema", puis quelques articles d'actualité le concernant.

En effet, en dépit de l'ambiguïté de la requête "Karim Benzema" (i.e., ne contenant aucune information temporelle), un modèle de RI typique devrait être en mesure d'identifier les documents qui pourraient être pertinents quant au critère thématique aussi bien que la dimension temporelle. Après une analyse temporelle de la requête, Google a conclu que la requête ne présente pas une spécificité temporelle, et donc a retourné des documents ne dépendant pas fortement du temps. Certains systèmes de RI retournent les documents dans l'ordre de leur création ou demandent tout simplement aux utilisateurs de spécifier la période de temps qui correspond à leurs intentions de recherche. Cependant, certaines requêtes pourraient concerner même des données liés au futur (e.g., nouveaux produits commerciaux attendus, une future compétition sportive, etc). Comprendre et analyser la

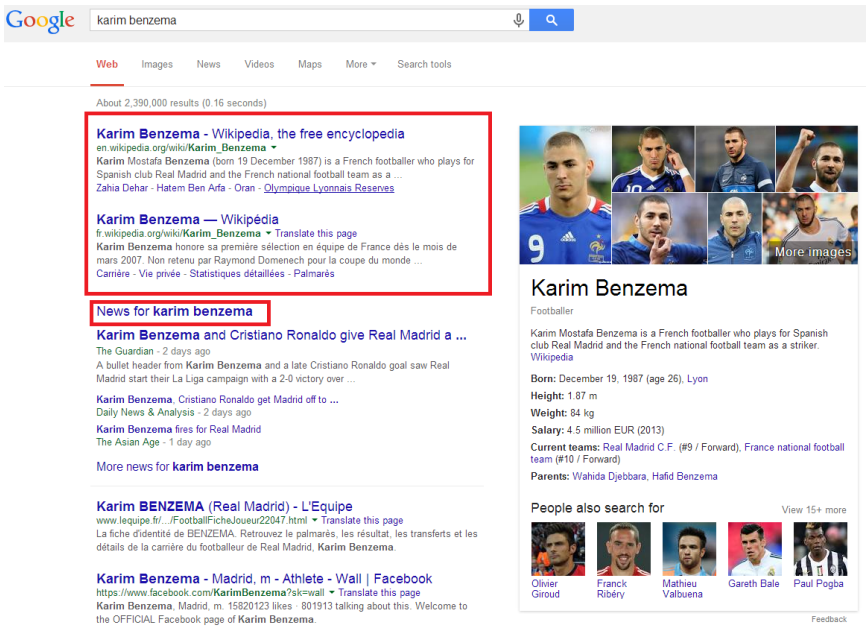


FIGURE 4.3: Résultats de recherche sur Google pour la requête "Karim Benzema", soumise le 29/06/2014.

dynamique des requêtes pourrait être une des solutions phares pour ces problématiques (Jones et Diaz, 2007; Metzler *et al.*, 2009; Kulkarni *et al.*, 2011; Radinsky *et al.*, 2012). Nous présentons les travaux qui ont été effectués sur deux types de requêtes :

Requêtes orientées récence : sont les requêtes qui surviennent juste après les toutes dernières nouvelles ou les événements plus récents (e.g., "grue mecque", "explosion Tianjin", etc). Le travail de Li et Croft (2003) est parmi les premières recherches abordant ce type de requêtes. Les auteurs ont classé les requêtes en fonction de la distribution temporelle des documents sur une collection de requêtes de TREC (volume 4 et 5). Dakka *et al.* (2012) ont défini les requêtes sensibles au temps comme : "*les requêtes pour lesquelles les documents pertinents ne sont pas distribués d'une façon uniforme au cours du temps, mais qui sont plutôt condensés dans des intervalles de temps restreints*". Si la période pertinente pour une requête sensible au temps ne peut pas être déterminée, les auteurs suggèrent le calcul d'une probabilité $p(q|t)$ pour chaque instant t et requête q en utilisant le modèle de vraisemblance.

Requêtes périodiques et bursts (rafales) : sont les requêtes qui sont soumises d'une façon récurrente dans les mêmes périodes de temps (e.g., "élection", "festival de cannes", etc). Vlachos *et al.* (2004) ont représenté ces classes de requêtes comme une transformation de Fourrier en se basant sur des séries chronologiques. Ces séries sont construites pour chaque terme de la requête à partir d'une large collection d'archives de requêtes du moteur de recherche MSN. Les périodes de temps importantes sont ensuite identifiées à travers une analyse exponentielle. Les périodes significatives sont celles présentant des lois de puissance différentes de la majorité des lois des autres périodes. Après l'extraction des périodes importantes, les auteurs ont proposé une technique d'identification de "bursts" basée sur le calcul de la moyenne mobile (*Moving Average*). Subasic et Castillo (2010) définit un "burst" de requête comme : "une période présentant un grand intérêt des utilisateurs sur un sujet qui a suscité une fréquence très élevée de requêtes qui lui sont liées". Subasic et Castillo (2010) a regroupé les requêtes bursts en trois classes : *(i)* celles qui disparaissent complètement après une certaine période ; *(ii)* les bursts sur des sujets existants ; et *(iii)* les bursts qui créent d'autres sujets. Dans la même direction de recherche, d'autres travaux ont souligné l'importance des descripteurs des requêtes dans des applications telles que le suivi d'événements et l'identification des intentions temporelles de recherche (Ginsberg *et al.*, 2009; Radinsky *et al.*, 2012, 2013; Joho *et al.*, 2014; Asur et Buehrer, 2009; Ren *et al.*, 2013; Costa *et al.*, 2014).

Prédiction et suivi des événements. Ginsberg *et al.* (2009) ont montré que les requêtes Google sont une source proéminente pour le suivi des maladies tels que l'*influenza*. Les auteurs ont calculé des séries chronologiques d'environ 50 millions des requêtes les plus similaires aux États Unies et ont proposé un modèle qui permet de détecter les requêtes liées aux maladies en relation avec l'*influenza*. Dans la même direction de recherche, Diaz (2009) a étudié les requêtes qui traitent des sujets liés aux actualités. Diaz a développé une méthode permettant, selon le type de requête, d'injecter des documents liés aux dernières nouvelles sur la requête. La figure 4.4 montre un exemple donné par Diaz, dans lequel ils a placé trois articles d'actualité liés à la requête "zimbabwe elections" (soumise en 2009). Cette technique est maintenant adoptée dans la plupart des méthodes de recherche, comme nous l'avons déjà montré.

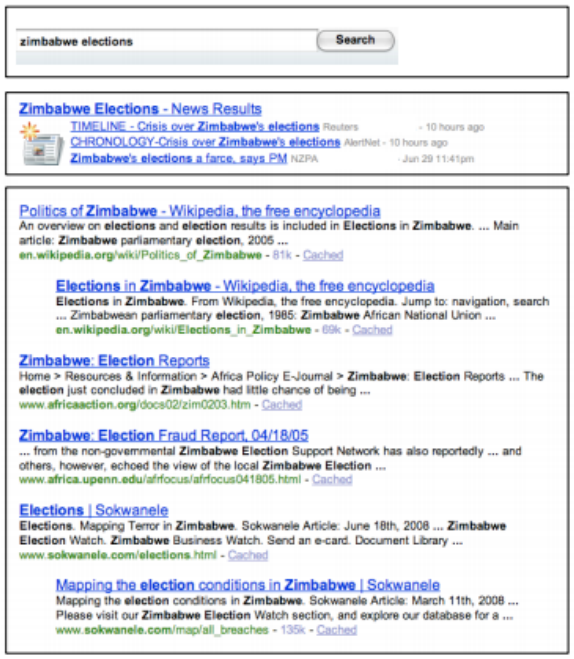


FIGURE 4.4: Exemple d'injection d'articles d'actualités en réponse à la requête "*zimbabwe elections*", soumise le 29/06/2009 (Diaz, 2009).

4.3.3 Le temps au niveau du contenu des documents

Dans cette section, nous présentons les études qui ont exploité les descripteurs temporels au niveau du contenu du document. Le défi principal ici consiste en l'identification et l'extraction des expressions temporelles contenues dans les documents. L'extraction de ces expressions est une étape cruciale dans la détermination des périodes de temps auxquelles fait référence un document. Ce problème fait partie de la tâche des systèmes d'annotation temporelle (*temporal taggers*). La première étape d'un système d'annotation consiste à segmenter le texte du documents en un ensemble de phrases. Dans la deuxième et troisième étape, le système identifie les phrases et effectue un étiquetage morpho-syntaxique puis essaye de reconnaître les entités contenues dans le texte. Une fois les expressions temporelles extraites, ils doivent être normalisées. Un des systèmes d'annotations temporelles les plus utilisés est HeidelTime [2]. HeidelTime est un système multilingue qui permet d'iden-

2. `http://heideltime.ifi.uni-heidelberg.de/heideltime/`

tifier et extraire les expressions temporelles de référence (Strötgen *et al.*, 2012). La figure 4.5 donne un exemple d'extraction à partir d'un document extrait d'un journal en ligne. HeidelTime annote les expressions en utilisant le standard d'annotation TIMEX [3].

```
The co-founder and former chief executive officer of Apple Inc. Steve Jobs has
died <TIMEX3 tid="t5" type="DATE" value="2011-10-05">yesterday</TIMEX3> at the
age of 56 , according to the company website .
On <TIMEX3 tid="t7" type="DATE" value="2011-08-24">August 24</TIMEX3> , Jobs
resigned from his post as CEO .
<s id="4"> He has been fighting pancreatic cancer since <TIMEX3 tid="t8" type=
"DATE" value="2004">2004</TIMEX3>
```

FIGURE 4.5: Exemple d'extraction d'expressions temporelles avec l'outil HeidelTime sur un extrait de document issu d'un journal du web.

De nombreux autres outils d'extraction d'expressions temporelles ont été aussi proposés. Le tableau 4.1 présente quelques outils avec la catégorie des expressions qu'ils sont capables d'extraire, s'ils effectuent une normalisation ou non et s'ils sont disponibles ou commerciaux.

Dans le même contexte, Lin *et al.* (2014b) ont utilisé l'outil GUTime [6] pour extraire les expressions temporelles et déterminer le temps référent des documents. Ce problème a fait aussi l'objet de plusieurs tâches d'évaluation tels que SemEval, TempEval, etc (Verhagen *et al.*, 2010). L'objectif de ces tâches est de proposer de nouvelles méthodes pour l'identification automatique de toutes les expressions temporelles, les événements et les relations temporelles dans les documents. Pour répondre à ces tâches, plusieurs approches basées sur le traitement automatique de langue ont été appliquées avec succès (Pustejovsky et Verhagen, 2009; UzZaman *et al.*, 2012).

4.3.4 Le temps au niveau des modèles d'ordonnancement

Dans cette section, nous montrons comment le temps est exploité au niveau des modèles d'ordonnancement dans des collections de documents qui changent dans le temps. La figure 4.6 donne un aperçu de l'architecture générale d'une approche de RI sensible au temps.

3. http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/time/Timex.html

6. http://timeml.org/site/tarsqi/modules/gutime/

	Catégorie	Normalisation	Disponibilité
HeidelTime (Strötgen *et al.*, 2012)	Explicite (E), Implicite (I), Relative (R)	Oui	Oui
GUTime [4]	E, I, R	Oui	Oui
SUTime (Chang et Manning, 2012)	E, I, R	Oui	Oui
Clinical TERN [5]	E, I, R	Oui	Oui
ManTIME (Filannino *et al.*, 2013)	E, I, R	Oui	Oui
PorTexto Craveiro *et al.* (2009)	E, R	Oui	Non
ANNIE Cunningham *et al.* (2002)	E, I, R	Non	Oui

TABLE 4.1: Outils d'extraction d'expressions temporelles.

Le modèle de langue sensible au temps est l'un des approches les plus utilisées pour la combinaison des critères temporels et thématiques (Li et Croft, 2003; Berberich *et al.*, 2010). Ce modèle utilise le temps pour donner plus d'importance aux documents récemment publiés. Li et Croft (2003) ont aussi défini un modèle qui utilise une distribution exponentielle pour répondre aux requêtes dont le besoin est purement temporel. Les documents les plus récents se sont attribués les scores les plus élevés :

$$p(d|T_d) = p(T_d) = \lambda e^{-\lambda(T_C - T_D)} \tag{4.1}$$

Où T_C est la date la plus récente dans toute la collection et T_D est la date de création du document. L'évaluation expérimentale dans la collection des requêtes TREC 301 – 400 (volume 4 et 5) montre que le modèles temporels sont meilleurs que les modèles de langue classique et de combinaison linéaire pour les requêtes dépendantes du temps.

Dans la même ligne de recherche, Dakka *et al.* (2012) ont modélisé la pertinence comme une combinaison de la dimension thématique et temporelle

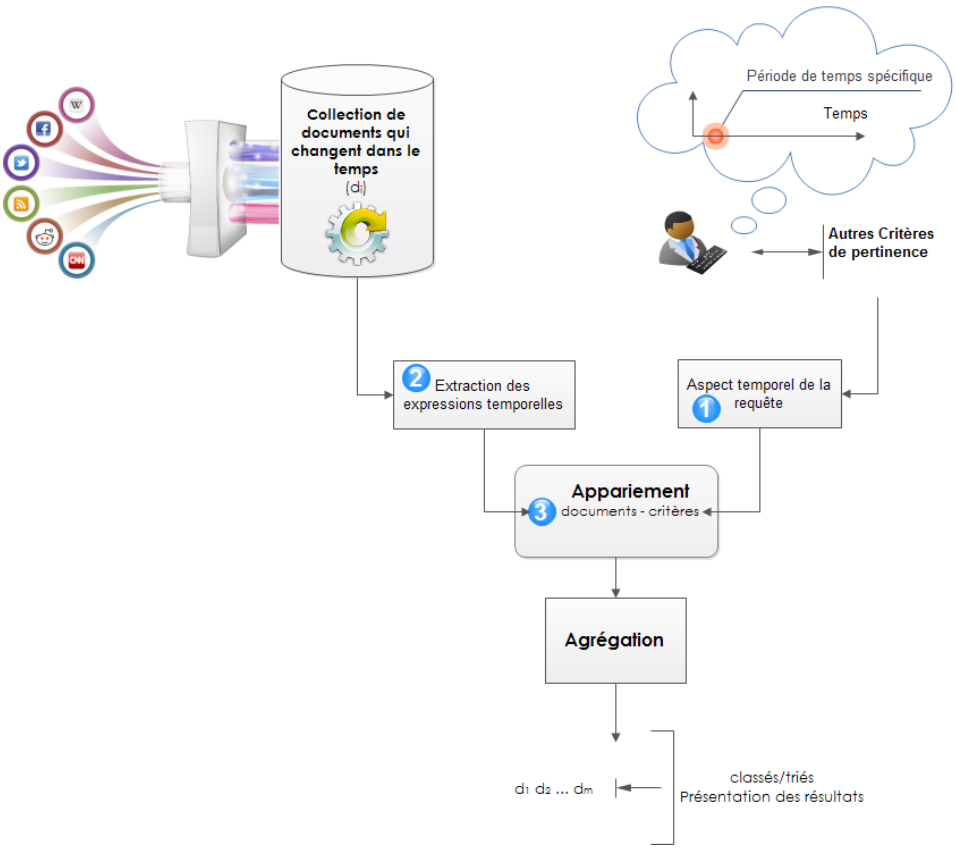


FIGURE 4.6: Processus général d'ordonnancement dans une approche de RI sensible au temps.

représentées respectivement par les probabilités $(P(q|d))$ et $(P(t|q))$. Le modèle identifie automatiquement les intervalles de temps importants pour la requête sensible au temps et les utilise pour favoriser les documents qui sont publiés dans ces périodes. La pertinence totale est donnée par :

$$P(d_t|q) = P(d,t|q) \propto P(q|d)P(t|q) \tag{4.2}$$

où $P(q|d)$ correspond au modèle de vraisemblance sur le document d et $P(t|q)$ représente l'importance relative du temps t pour la requête q. Le facteur temporel $P(t|q)$ peut être estimé en utilisant le maximum de vraisemblance, défini comme la somme normalisée des scores de pertinence des documents publiés à l'instant t.

Lin *et al.* (2012) ont proposé un moteur de recherche sensible au temps

(TASE : Time-Aware Search Engine) se basant sur la fréquence des informations temporelles et la relation entre elles. Comme pour la plupart des approches temporelles, le système TASE combine les scores de similarité textuelles et temporelles. La méthode de combinaison est basée sur un modèle linéaire qui s'est inspiré de quelques travaux de la littérature (Kanhabua et Nørvåg, 2011). La figure 4.7 montre un exemple de recherche sur TASE avec la requête "Michael Jackson".

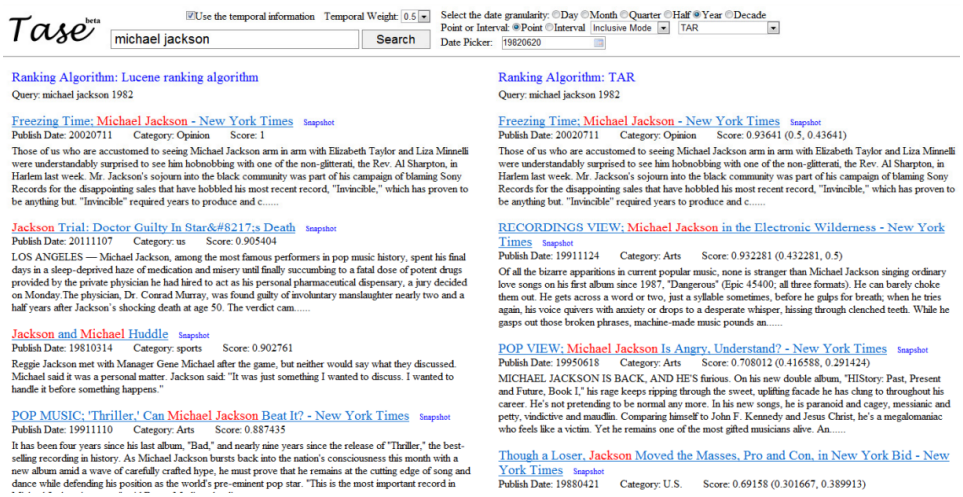


FIGURE 4.7: Exemple de recherche avec le modèle de recherche temporel TASE (Lin *et al.*, 2012).

Plusieurs autres modèles de recherche temporels ont été proposés. M. *et al.* (2010) ont proposé *Time Explorer*, un outil de recherche qui donne des résultats sous forme de *timelines*. Ce modèle permet l'analyse des changements de topics dans le temps au sein d'une collection d'archives d'actualités. Campos *et al.* (2014b) ont défini un nouveau modèle d'ordonnancement sensible au temps appelé GTE-Rank considérant l'importance thématique et la distance temporelle pour le ré-ordonnancement des snippets Web. Le modèle, déployé en ligne [7], permet la recherche de topics liés au temps en les classant sous forme de *timeline*. Harper et Chen (2012) assume que le Web est en évolution et suppose que ce changement n'affecte pas la similarité textuelle des pages Web aux requêtes, contrairement à leur importance qui change dans le temps. Les pages sont classées en fonction de leur contenu, l'information temporelle qu'ils incluent et leur importance dans le temps. Les auteurs ont

7. `http://www.ccc.ipt.pt/~ricardo/software.html`

appliqué l'algorithme PageRank sur ces trois facteurs et ont montré leur efficacité dans la recherche de documents pertinents. Dans la même ligne de recherche, Perkiö *et al.* (2005) ont montré que la pertinence n'est pas statique dans le temps et ont proposé un modèle statistique se basant sur les caractéristiques temporels des descripteurs des documents. L'hypothèse des auteurs est que l'ordonnancement des résultats pour une requête q à un instant t doit favoriser les documents dont les sujets importants sont les mêmes que la plupart de ceux qui sont actifs dans toute la collection à l'instant t. Cette hypothèse est modélisé à travers une adaptation du modèle de pondération $TF-IDF$. Cette adaptation temporelle est est obtenue à travers la formule suivante :

$$score(d,q;t) = \frac{s(d,q)}{\sum_{k\epsilon\mathcal{K}(t)} log1/c_k^{M(t)} + log1/c_k^d} \tag{4.3}$$

Où $s(d,q)$ est le score du document d suivant le modèle, $TF-IDF$ $\mathcal{K}(t)$ est l'ensemble des indices des K topics les plus important dans le modèle à l'instant t. $c_k^{M(t)}$ représente l'importance du topic k dans le modèle à l'instant t, et c_k^d est l'importance du topic k dans le document d.

Les expérimentations effectuées avec ce modèle sensible au temps ont montré des résultats encourageants. Plusieurs autres travaux ont aussi confirmé cette hypothèse (Efron, 2010; Karkali *et al.*, 2014; Aji *et al.*, 2010; Nunes *et al.*, 2011; Wang *et al.*, 2014; Kulkarni *et al.*, 2011; Kim *et al.*, 2013).

Efron (2010) a utilisé la distribution temporelle des termes des requêtes dans la collection entière, plutôt que dans les documents, pour dériver le poids de chaque mot à utiliser dans l'ordonnancement. Le poids total d'un terme t étant donnée le modèle M de la série temporelle pour ce dernier est donné par :

$$poids_M(t) = \sqrt{\frac{\sum_{i=1}^{n}(\hat{t}_i - t_i)^2}{n-2}} \tag{4.4}$$

Où $\hat{t}_i$ est le modèle estimé de t_i, et n est la longueur de la série temporelle observée pour t. Contrairement aux mesures d'importance des termes traditionnelles, pour lesquelles les termes rares reçoivent plus de poids que les termes fréquents, Efron a défini un schéma de pondération temporel basé sur les séries chronologiques. Ce modèle a donné des bonnes performances comparativement aux méthodes classiques, sur des données de TREC. Aji

et al. (2010) ont aussi proposé une méthode similaire qui est basé sur l'historique de modification des documents. L'intuition derrière cette méthode est que l'importance d'un terme peut être mesurée par l'analyse de l'historique de révision d'une page qui contient des informations précieuses des connaissances fournies par les éditeurs (e.g., historique d'édition des pages Wikipedia). Comme le contenu des documents peut changer au cours du temps à cause des révisions reflétant des événements liés aux topics, des termes important peuvent êtres ajoutés aux documents. Le nouveau poids dépendant du temps d'un terme t est est défini par :

$$poids_{global}(t,d) = \sum_{j=1}^{n} \frac{c(t,v_j)}{j^{\alpha}} \tag{4.5}$$

Où n est le nombre de révisions effectuées sur le document d. La fréquence brute du terme t dans la révision j est dénotée par $c(t,v_j)$, et elle est modifiée en utilisant le facteur de décroissance temporelle (*decay factor*) j^{α}, où α est un paramètre qui contrôle la vitesse de cette décroissance. Ainsi, j^{α} permet d'ajuster l'importance relative d'un terme sur les différentes révisions du document. Par exemple, dans les révisions antérieures de la page Wikipedia du film *Avatar* (Juin 2006), il n'y avait que quelques modifications dans le titre et le contenu qui ont simplement mentionné que *James Cameron* aurait dirigé le film. Cependant, en Octobre 2006, des changements majeurs sont survenues sur le contenu des la page, comme des détails sur le budget et le déroulement ont été ajoutés. Les auteurs assument que l'importance du terme doit être ajustée en intégrant l'historique des *bursts*. Cette observation a donné lieu à un modèle qui combine l'importance des termes calculée à partir des historiques de révisions des pages Web avec des méthodes statistiques tels que $BM25$ et le modèle de langue. L'évaluation expérimentale dans le cadre de la tâche ad-hoc sur la collection INEX 2009 montre que le modèle proposé est plus performant que les méthodes classiques. Toutefois, ce modèle n'est pas performant pour les requêtes ambiguës dont les documents pertinents peuvent traiter plusieurs autres topics.

4.3.5 Synthèse

Les recherche en RI sensible au temps ont montré que la dimension temporelle est primordiale dans le processus d'ordonnancement des documents (Yu *et al.*, 2004; Nunes, 2007). Plusieurs sources d'évidence sont exploitées comprenant des descripteurs liés aux documents et aux requêtes pour l'amélioration des performances des SRI. nous avons classé ces méthodes en trois

catégories suivant le niveau dans lequel le temps est exploité (document, requête et modèle). Le tableau 4.2 donne un aperçu des travaux effectués dans ce cadre, avec la classe de requête, le modèle d'ordonnancement et la collection de test utilisée. Les requêtes étudiées sont celles pour lesquelles le besoin utilisateur concerne des documents publiés dans des périodes de temps spécifiques dans le passé ou des documents qui sont récemment publiés. Les études concernant l'élicitation du type des requêtes sont souvent basées sur l'exploitation de la distribution des documents pertinents pour identifier les intervalles de temps pertinents pour la requête. Considérons par exemple la requête "explosion", les pics dans la distribution des documents peuvent indiquer les périodes dans lesquelles il y a eu une explosion. Les séries chronologiques ont été largement exploitées pour ce genre de problèmes surtout dans les collections de documents qui changent dans le temps et qui sont liées aux actualités (Vlachos *et al.*, 2004; Radinsky *et al.*, 2013; Kim *et al.*, 2013; Shokouhi, 2011). Cette particularité peut aider les moteurs de recherche à choisir les périodes dans lesquelles il convient d'injecter des actualités dans les résultats de recherche.

Du point de vue d'ordonnancement, de nombreuses études exploitent le changement du comportement de recherche des utilisateurs et des collections pour améliorer les résultats de recherche (Radinsky *et al.*, 2012). Tandis que des approches ont exploré des requêtes contenant des indicateurs temporels explicites (e.g., années) pour modifier le modèle de langue (Metzler *et al.*, 2009), d'autres ont ajouté des critères temporels dans le même modèle pour re-ordonner les résultats (Li et Croft, 2003). Le changement de la fréquence des termes durant le temps a été aussi étudié pour la pondération des termes dans les collections dynamiques (Efron, 2010). Le modèle de langue temporel a été parmi les modèles les plus exploités dans divers application de RI temporelle (Li et Croft, 2003; Berberich *et al.*, 2010; Dakka *et al.*, 2012).

Travaux de recherche	**Type de requêtes**	**Modèle d'ordonnancement**	**Collection de données**
Li et Croft (2003)	Orientées récence	Modèle de langue sensible au temps	requête 301 – 400 des collections TREC (volumes 4 et 5)
Vlachos *et al.* (2004)	Bursts, périodiques	Séries chronologiques	Logs de requêtes du moteur de recherche MSN
Jones et Diaz (2007)	Atemporelle, temporellement ambiguës, temporellement non ambiguës	-	Articles d'actualités de logs de recherche
Diaz (2009)	-	Modèle de langue	Articles d'actualités
Metzler *et al.* (2009)	Requête implicitement contenant des dates (années)	-	logs de requêtes
Asur et Buehrer (2009)	Navigationnelle, adulte et requêtes sur les actualités	-	-
Berberich *et al.* (2010)	Expressions temporelles	Modèle de langue sensible au temps	Le corpus annoté de "New York Times", Wikipedia
Subasic et Castillo (2010)	Bursts	Modèle de langue sensible au temps	Logs de requêtes Yahoo !
Shokouhi (2011)	Saisonnières	-	Logs de recherche
Kulkarni *et al.* (2011)	Periodicité, tendances (trends)	-	Logs de requête Bing, archives, jugement de pertinence périodiques
Radinsky *et al.* (2012)	Tendances, periodicité, bruit, surprise, Saisonnières	Séries chronologiques	Logs de requêtes et URL de Bing
Dakka *et al.* (2012)	Sensibles au temps	Modèle de langue temporel	Articles d'actualités
Efron (2010)	-	Séries chronologiques	TREC Tipster, tâche TREC robust

TABLE 4.2: Une synthèse de quelques travaux sur la RI sensible au temps.

4.4 Évaluation des méthodes de recherche d'information temporelle

Dans cette section, nous présentons quelques tâches de recherche permettant l'évaluation des modèles de RI sensibles au temps. Motivé par l'intérêt théorique d'évaluer ces modèles, de nombreux compagnes d'évaluations ont été proposées. Ces compagnes incluent, mais ne sont pas limitées à :

SemEval 2015 - Tâche 4 (*Time and Space track*[8]) : Cette tâche d'évaluation inclu 5 tâches : *(i) TimeLine : Cross-Document Event Ordering* ; *(ii) QA TempEval* ; *(iii) Clinical TempEval* ; *(iv) Diachronic Text Evaluation* ; et *(v) SpaceEval*. La première tâche *Cross-Document Event Ordering* s'intéresse au problème de génration de *timeline* à partir des collections de données liées à des entités (i.e., organisations, personnes, événements, etc). La tâche *QA TempEval* traite le problème de questions-réponses d'un point de vue temporel. L'objectif consiste à construire une base connaissances pour répondre à des questions à caractère temporel et les comparer à des réponses données par des utilisateurs réels. La tâche *Clinical TempEval* vise à identifier et décrire des événements et les relations entre eux dans textes médicaux. L'identification automatique des périodes de temps dans lesquelles les événements sont écrits est traités dans la tâche *Diachronic Text Evaluation*. La tâche *SpaceEval* consiste à identifier et classer les items d'un point de vue spatial en considérant conjointement les critères géographiques et temporels.

Tâche TREC Temporal Summarization[9] (TS) : a pour objectif de retourner des documents pertinents pour les événements à partir d'une collection dont tous les documents sont estampillés. Cette tâche comprend deux sous tâches : *(i) Sequential Update Summarization*, où les participants doivent retourner des mises à jour pertinentes pour un événement ; et *(ii) Value Tracking*, qui consiste estimer la valeur des attributs pour un événement (eg., nombre de blessés suite à un séisme). Les participants à cette tâche utilisent la collection de données fournie par la tâche KBA de TREC. 8 groupes ont participé à TREC TS 2013, le meilleur système a été proposé par le groupe PRIS (Zhang *et al.*, 2013). Pour la première tâche, le système proposé s'est basé sur le modèle hiérarchique LDA, alors qu'il a exploité le modèle *Conditional Random Field (CRF)* pour la seconde tâche.

Tâche TREC Knowledge Base Acceleration (KBA) (Frank *et al.*, 2012) : la

8. `http://alt.qcri.org/semeval2015/task4/`
9. `http://www.trec-ts.org/`

tâche KBA a pour objectif de guider les utilisateurs à maintenir les grandes bases de connaissances telles que Wikipedia, en filtrant automatiquement les documents qui sont potentiellement pertinents vis-à-vis d'une liste prédéfinie d'entités. Dans la tâche TREC KBA 2012, les organisateurs ont sélectionné un ensemble de topics en se basant sur 29 entités extrait à partir de Wikipedia : 27 personnes et 2 organisations. Les participants (Abbes *et al.*, 2013; Bouvier et Bellot, 2014) ont appliqué leurs modèles sur un corpus organisé sous forme de répertoires, où chaque répertoire correspond à une heure, pour simuler le flux de données en temps réel. Le meilleur système dans KBA 2013 (11 groupes participants, 43 systèmes) s'est basé sur des techniques d'apprentissage d'ordonnancement (eg., SVM). La tâche TREC KBA 2013 a proposé deux sous tâches : *(i)* Cumulative Citation Recommendation (CCR), et *(ii) Streaming Slot Filtering (SSF)*. 141 entités ont été fournies et 15 groupes ont participé à la tâche avec 140 systèmes soumis. Le meilleur système s'est aussi basé sur une méthode d'apprentissage automatique.

Tâche NTCIR Temporal Information Access (Temporalia)[10] : propose deux sous tâches pour traiter les problèmes de RI sensibles au temps : *(i) Temporal Query Intent Classification (TQIC)* ; et *(ii) Temporal Information Retrieval (TIR)*. Tandis que TQIC s'intéresse à la classification des requêtes en 4 classes temporelles prédéfinies (passé, future, atemporelle, récente) en se basant sur l'intention implicite ou explicite, TIR s'intéresse à la recherche de documents en réponse à des topics qui intègrent le temps comme facteur de pertinence (Joho *et al.*, 2014).

Tâche Tweet timeline generation (TTG) (Lin *et al.*, 2014a) : est une nouvelle sous tâche de TREC Microblog 2014 dont l'objectif est répondre à la problématique suivante : "on a un besoin en information décrit par une requête q soumise à l'instant t, on souhaite avoir un résumé qui décrit brièvement la requête q". Deux défis à relever dans ce contexte : *(i)* détecter et éliminer les tweets redondants (nouveauté) ; et *(ii)* automatiquement identifier le nombre de tweets pertinents à retourner. Dans TREC TTG 2014, 15 groupes ont participé à la tâche avec 50 systèmes. Le meilleur système dans cette tâche est basé sur un graphe qui représente les tweets comme des noeuds et utilise des mesures de similarité entre le tweets d'une part, et entre les tweets et les requêtes d'une autre part (Zhang *et al.*, 2014).

Nous présentons dans le tableau 4.3, une synthèse sur les collections de données utilisées, les fenêtres temporelles de chacune ainsi que les mesures

10. `https://sites.google.com/site/ntcirtemporalia/`

d'évaluation utilisées.

Tâche	Collection de données	Fenêtre temporelle	Disponibilité	Mesures d'évaluation
SemEval 2015	Collection d'articles d'actualités, wiki, blogs, données cliniques	différentes fenêtres (du 1960 à 2014)	■	MAP, rappel, précision
TREC TS	Corpus TREC KBA	Octobre 2011 jusqu'à mi-février 2013	■	EG, *coverage*, EL, Moyenne harmonique de EL normalisée, *Latency Comprehensiveness*
TREC KBA	Corpus TREC KBA	Octobre 2011 jusqu'à mi-février 2013	■	F_1 score, *Scaled Utility*
NTCIR Temporalia	articles LivingKnowledge	2014	■	Précision, nDCG, Q-measure
TREC Microblog TTG	corpus TREC Microblog	2014	■	*Cluster precision*, *Cluster recall*

TABLE 4.3: Tâche d'évaluation des modèles de RI sensibles au temps.

4.5 Conclusion

Nous avons présenté dans ce chapitre les travaux de la littérature exploitant le temps dans les requêtes, documents et modèle d'ordonnancement ainsi que les principales méthodes de combinaison du critère temporel avec les autres descripteurs de pertinence. Dans un premier temps, nous avons souligné l'importance de l'étude des caractéristiques temporelles des requêtes et l'impact de ces caractéristiques dans le processus d'ordonnancement. Plusieurs techniques d'élicitation des types de requêtes sont également proposées. Au niveau document, nous avons présenté les méthodes d'extraction des expressions temporelles. Enfin, nous avons abordé le problème d'ordonnancement sensible au temps en présentant les principales techniques liées

à la combinaison des scores temporels et thématiques. La particularité des approches présentés dans ce chapitre et la nature des collections de données exploitées qui ne sont plus statique, tel est le cas dans le chapitre 3. Les collections et donc le contenu des documents changent dans le temps, ce qui explique que la plupart des méthodes ont été testées sur des logs de requêtes de moteurs de recherche existant ou des articles d'actualités et Wikipedia pour bien refléter la réalité.

Deuxième partie

Contribution à la définition et l'évaluation de modèles d'agrégation de pertinence multidimensionnelle en RI

Chapitre 5

Méthode d'agrégation de pertinence multidimensionnelle : proposition et évaluation dans des tâches de RI sociales et personnalisées

---

5.1 Introduction

Dans ce chapitre, nous proposons une approche basée sur l'intégrale de Choquet discrète pour l'agrégation de pertinence multidimensionnelle. Fondée sur les intégrales floues et basée sur le concept de la mesure floue, la principale originalité de l'intégrale de Choquet, en plus de sa généralisation des opérateurs d'agrégation classiques tels que les moyennes arithmétiques et ses variantes, réside dans sa capacité à modéliser toutes les interactions et les dépendances qui peuvent exister entre les différentes dimensions de pertinence. Nous tentons de tenir compte de la propriété de subjectivité qui peut se décliner à travers les différences entre les utilisateurs quant à l'importance accordée à chaque dimension de pertinence. Nous nous basons sur la flexibilité de la mesure floue qui est à la base de la quantification de l'importance estimée de chaque dimension pour chaque utilisateur ainsi que

leur degré d'interaction ou d'interdépendance. Ainsi, nous proposons deux modèles d'agrégation de pertinence multicritères :

- Une approche multicritère générique basée sur l'intégrale de Choquet pour l'agrégation de pertinence multidimensionnelle (Moulahi *et al.*, 2013, 2014d). Nous proposons également un nouvel algorithme d'apprentissage des mesures d'importance des critères. Cette approche est évaluée dans une tâche spécifique de recherche de tweets, où les critères conjointement considérés sont : la dimension thématique de recherche, la fraîcheur des tweets et l'autorité des utilisateurs. Les expérimentations ont été menées sur la collection de test fournie par la tâche Microblog de TREC 2011 et 2012.
- Une approche multicritère personnalisée qui se base sur le même opérateur de Choquet (Moulahi *et al.*, 2014b,c), mais qui s'en distingue selon les points clés suivants :

  1. Une agrégation pondérée par les préférences des utilisateurs quant à chacune des dimensions agrégées, proposant de déployer un mécanisme d'agrégation produisant des scores de pertinence dépendant des préférences des utilisateurs ;

  2. Une nouvelle évaluation expérimentale tant dans l'objectif que dans la méthodologie (Moulahi *et al.*, 2014a), en utilisant la collection de test standard TREC *Contextual Suggestion*. Nous montrons l'impact de la prise en compte des dépendances entre les critères ainsi que leur personnalisation sur les performances de recherche.

La suite du chapitre est organisée comme suit : dans la section 5.2, nous donnons la problématique et quelques motivations, et nous formalisons le problème d'agrégation de pertinence multidimensionnelle. Ensuite, nous présentons les spécificités de notre opérateur d'agrégation généralisé pour la RI dans les sections 5.3 et 5.4. La section 5.5 détaille les principes d'agrégation personnalisée ainsi que la méthode d'apprentissage des mesures d'importance pour chaque utilisateur. L'évaluation des méthodes proposées est détaillée dans la section 5.6. Dans la section 5.6.2.1, nous présentons l'évaluation expérimentale de notre première approche dans un contexte de RI sociale. Nous définissons ainsi le cadre expérimental et les résultats obtenus. Les sections 5.6.4 et 5.6.5 décrivent respectivement le cadre expérimental puis les résultats de l'application de l'approche proposée dans la tâche TREC dédiée à la RI personnalisée en l'occurrence "TREC Contextual Suggestion" (Dean-Hall *et al.*, 2013) et une tâche de RI personnalisée dans les folksonomies (Vallet et Castells, 2012).

5.2 Formalisation du problème et positionnement

5.2.1 Formalisation du problème

Les fonctions d'agrégation sont généralement définies et utilisées pour combiner plusieurs valeurs numériques ou préférences liées à des critères en une seule, de telle sorte que le résultat final de l'agrégation prenne en compte, d'une manière prescrite, toutes les valeurs individuelles. Comme nous l'avons déjà montré dans le chapitre 3, de nombreuses fonctions d'agrégation multicritères ont été proposées telles que la moyenne arithmétique, la méthode de combinaison linéaire, les méthodes d'apprentissage automatique et bien d'autres encore. Dans ce manuscrit, nous nous limiterons cependant aux fonctions d'agrégation qui associent une valeur numérique à chaque critère de pertinence, lesquels représentent des documents (ou des alternatives). Nous ne traiterons pas les fonctions de combinaison multicritères qui, de façon plus générale, permettent de ranger les documents sans leur assigner des valeurs précises. Ainsi par exemple, les méthodes de surclassement (ELECTRE, MACBETH, etc) sont des procédures de rangement, plutôt que des fonctions d'agrégation à proprement parler. Nous supposerons donc que les valeurs à agréger appartiennent à des échelles numériques, qui sont de type cardinal. Une fois que les valeurs à agréger sont définies, nous pouvons les combiner en une seule valeur au moyen de la fonction d'agrégation. Mais une telle opération, comme nous l'avons déjà mentionné, peut s'effectuer de nombreuses façons selon ce qui est attendu de la fonction d'agrégation, selon la nature des valeurs à agréger et selon la relation entre les critères à agréger.

D'une façon générale, le problème d'agrégation de pertinence multidimensionnelle peut être perçu comme un problème de prise de décision multicritère (Cf., chapitre 3). Plus précisément, dans ce contexte de RI, nous sommes confrontés à trouver un consensus sur l'ordonnancement d'un ensemble de documents $d_j \in \mathcal{D}$ selon un ensemble de critères de pertinence $c_i \in \mathcal{C}$, en réponse à une requête utilisateur. La combinaison des scores partiels des documents, obtenus sur chaque dimension de pertinence est généralement donnée par une fonction ayant la forme suivante :

$$\mathcal{F} : \left\{ \begin{array}{l} \mathbb{R}^N \longrightarrow \mathbb{R} \\ (C_{1j} \times C_{2j} \times \ldots \times C_{Nj}) \longrightarrow \mathcal{F}(C_{1j}, C_{2j}, \ldots, C_{Nj}) \end{array} \right\}$$

Où $\mathcal{F}$ est la fonction estimant les scores globaux des documents, et C_{ij} (ou $C_i(d_j)$) est le score partiel obtenu sur le document d_j selon le critère c_i. C_{ij} est interprété comme le degré de satisfaction du document d_j selon le critère c_i. Selon le domaine considéré, les valeurs C_{ij} peuvent aussi être appelées des *performances*, obtenues selon une fonction d'utilité ou un critère. En outre, les préférences sur les critères sont souvent exprimées selon un vecteur de poids normalisés $W = (w_l, w_2, \ldots, w_n)$ (avec $0 \leq w_i \leq 1$) ou par une relation de préférence binaire $\succ_C$. La relation $d_1 \succ_C d_2$ peut être interprétée par *"d_1 est plus pertinent que d_2 selon l'ensemble $\mathcal{C}$ des dimensions de pertinence"*. De la même façon, une relation de préférence partielle $\succ_{c_i}$ peut être introduite également sur l'ensemble des documents. Ainsi, $d_1 \succ_{c_i} d_2$ peut être vue comme : *"d_1 est plus pertinent que d_2 selon la dimension de pertinence c_i"*.

En effet, le choix de la fonction $\mathcal{F}$ et l'utilisation de telle ou telle fonction doit toujours être justifiée. Pour choisir une fonction d'agrégation appropriée à n'importe quel problème de combinaison multicritères, il est utile d'adopter une approche *axiomatique* et sélectionner ainsi les fonctions d'agrégation selon des propriétés qu'elles vérifient. Par exemple, dans notre problème d'agrégation multicritère, le problème majeur est d'estimer le score global d'un document à partir des scores partiels obtenus sur les différents critères. Dans ce cas, il ne serait pas très naturel de donner au score global une valeur inférieure au plus petit des scores partiels ou supérieure au plus grand des scores partiels. Ainsi, seule une fonction de type "interne" (e.g., une moyenne) peut être utilisée. Dans ce travail, l'objectif n'est de proposer une approche axiomatique en particulier, néanmoins nous présentons quelques propriétés qui améliorent notre compréhension des fonctions d'agrégation et qui peuvent être vues comme souhaitables dans le domaine de la RI multicritères. Dans la suite, nous présentons les propriétés mathématiques les plus utiles pour notre contexte d'agrégation.

Propriété 1. *(Continuité)* Un opérateur d'agrégation est dit continu si la fonction d'agrégation est continue dans le sens usuel de ce terme, c'est à dire que pour tout $n \in \mathbb{N}$, l'opération n-aire $\mathcal{F}$ est continue. Cette propriété est souvent nécessaire dans de nombreuses applications, puisque elle contraint l'opérateur à ne pas se comporter de manière chaotique (Bouyssou *et al.*, 2006). En effet, l'avantage d'une fonction d'agrégation continue est qu'elle ne présente aucun saut brusque suite à de faibles variations des valeurs partielles.

Propriété 2. *(Monotonie)* Une fonction d'agrégation $\mathcal{F}$ est dite *monotone*

(ou non décroissante) si :

$$\forall n \in \mathbb{N} : x_1 \leq y_1, \ldots, x_n \leq y_n \Rightarrow \mathcal{F}(x_1, \ldots, x_n) \leq \mathcal{F}(y_1, \ldots, y_n) \tag{5.1}$$

Où $X = (x_1, x_2, \ldots, x_n)$ est le vecteur des scores (ou performances). Cette propriété peut être interprétée par : *si le degré de satisfaction d'un critère augmente, alors le score d'agrégation global augmente en conséquence.*

$\mathcal{F}$ est dite *unanimement croissante* si elle est non décroissante et si, pour tous vecteurs x et $y \in \mathbb{R}^N$, on a :

$$x_1 < y_1, \ldots, x_n < y_n \Rightarrow \mathcal{F}(x_1, \ldots, x_n) < \mathcal{F}(y_1, \ldots, y_n) \tag{5.2}$$

Une fonction non décroissante présente un comportement non négatif à tout accroissement des valeurs. En d'autres termes, l'accroissement d'une valeur partielle ne fait pas décroître le résultat. La fonction est strictement croissante si, en plus, elle réagit positivement à tout accroissement d'au moins une valeur partielle. Enfin, la fonction est *unanimement croissante* si elle est non décroissante et présente une réaction positive chaque fois que tous les arguments croissent.

Propriété 3. *(Idempotance)*

En algèbre, l'idempotence est une propriété liée à une opération "$*$" par laquelle un élément x est *idempotent*, c'est à dire $x * x = x$. Dans le cadre des opérateurs d'agrégation n-aires, cette notion s'étend à :

$$\mathcal{F}(x, \ldots, x) = x \tag{5.3}$$

Dans ce cas, $\mathcal{F}$ est dite *idempotente* si la propriété 5.3 est respectée pour tout x dans l'ensemble des valeurs. Cette propriété, aussi appelée *unanimité*, s'interprète de la manière suivante : *si on agrège n fois la même valeur, alors on s'attend à obtenir cette même valeur initiale.* Une fonction qui vérifie à la fois la propriété d'idempotance ainsi que la monotonie est une méthode qui présente un comportement de compensation, et qui peut être restreinte dans ce cas aux opérateurs respectant : $min \leq \mathcal{F} \leq max$.

Propriété 4. *(Condition aux limites)* On dit que $\mathcal{F}$ satisfait les conditions aux bornes si :

$$\mathcal{F}(0, \ldots, 0) = 0 \tag{5.4}$$

$$\mathcal{F}(1, \ldots, 1) = 1 \tag{5.5}$$

C'est l'unanimité pour les valeurs extrêmes.

Propriété 5. Une fonction d'agrégation $\mathcal{F}$ est dite :
– *Conjonctive* si $\mathcal{F}(x) \leq min(x_i)$, pour tout $x \in \mathbb{R}^N$
– *Disjonctive* si $max(x_i) \leq \mathcal{F}(x)$, pour tout $x \in \mathbb{R}^N$
– *Interne* si $min(x_i) \leq \mathcal{F}(x) \leq max(x_i)$, pour tout $x \in \mathbb{R}^N$

Toutes ces caractérisations axiomatiques définissent des familles de fonctions d'agrégation différentes, donnant ainsi des applications différentes dans divers domaines. Chaque méthode présente des inconvénients et des avantages qui varient selon le domaine et selon le cadre d'application. Dans ce chapitre, nous nous intéressons uniquement aux méthodes s'appliquant à la RI multicritères. Nous présentons dans la section suivante, les limites des fonctions d'agrégation classiques qui sont majoritairement utilisées pour l'estimation de pertinence multidimensionnelle.

5.2.2 Limites des opérateurs d'agrégation classiques pour la modélisation de pertinence

Bien que la plupart des méthodes existantes pour l'agrégation de pertinence se basent sur des combinaisons linéaires, et ce pour des raisons évidentes de simplicité et de complexité, il est bien connu que celles ci présentent des défauts fondamentaux qu'il n'est pas possible d'éliminer. Nous montrons à travers l'exemple qui suit une des insuffisances des mécanismes d'agrégation basés sur la combinaison linéaire.

Exemple 1 *Considérons un problème d'ordonnancement avec deux critères (c_1, c_2) et trois documents d_1, d_2 et d_3 ayant les scores partiels suivants :*
$C_1(d_1) = 0.45$, $C_1(d_2) = 0$, $C_1(d_3) = 1$
$C_2(d_1) = 0.45$, $C_2(d_2) = 1$, $C_2(d_3) = 0$.

Supposons que ces scores varient entre 0 et 1 et que le document d_1 est préféré à d_2 et d_3 (i.e., $d_1 \succ d_2 \sim d_3$). Ceci dit, les documents qui sont jugés de façon balancée, en terme de score, sur les deux critères, sont préférés aux documents satisfaisant un seul critère. La question majeure qui se pose ici est : comment pourrait-on modéliser ce type de préférence avec un opérateur d'agrégation de type moyenne ou une méthode de combinaison linéaire ?

Un des aspects significatifs dans ces méthodes d'agrégation est la prise en compte de l'importance des critères, laquelle est habituellement modélisée par l'utilisation de poids. Pour ce faire, il serait nécessaire de trouver les poids w_1 et w_2 de c_1 et c_2 respectivement, de telle sorte que les préférences

puissent être modélisées par ces opérateurs d'agrégation. Nous aurons donc :

$d_2 \sim d_3 \Leftrightarrow w_1(C_1(d_2)) + w_2(C_2(d_2)) = w_1(C_1(d_2)) + w_2(C_2(d_3)) \Rightarrow w_1 = w_2$
$d_1 \succ d_2 \Leftrightarrow w_1(C_1(d_1)) + w_2(C_2(d_1)) > w_1(C_1(d_2)) + w_2(C_2(d_2))$
$\Rightarrow 0.45 \times (w_1 + w_2) > w_2$

Par conséquent, nous obtenons : $0.9 \times w_2 > w_2$*, ce qui est absurde !*
En effet, aucune méthode d'agrégation additive ne permet de modéliser cette préférence. Il est bien connu dans le domaine d'agrégation multicritères que ces fonctions conduisent à l'indépendance préférentielle mutuelle parmi les critères, qui exprime, dans un certain sens, l'indépendance des critères. Comme ces fonctions ne sont pas appropriées en présence d'attributs dépendants, assez souvent, la tendance a été de construire des attributs censés être indépendants.

Pour pallier à cet inconvénient, des opérateurs d'agrégation prioritaires (Cf., Chapitre 3, Section 3.3.3) ont été récemment proposés (da Costa Pereira *et al.*, 2009, 2012). Néanmoins, malgré leur efficacité dans plusieurs cadres de RI (Bouidghaghen *et al.*, 2011b; da Costa Pereira *et al.*, 2012), il existe toujours des préférences qui ne peuvent pas être représentées par ces derniers. Formellement, le processus de priorisation est considéré uniquement sur les critères individuels (*e.g.,* $c_i \succ c_j \succ c_k$). Si on considère, par exemple, quatre critères : *aboutness* (Ab), fraîcheur d'information (Fr), *centres d'intérêt* (Ci) et autorité (Au), et on considère un utilisateur préférant les documents récents traitant un *topic* donné, plutôt que les documents *autoritaires* représentant au mieux le sujet recherché par l'utilisateur. Ainsi, on aura la préférence $\{Ab, Fr\} \succ_{\mathcal{C}} \{Au, Ci\}$, qui ne peut pas être modélisée par les opérateurs d'agrégation ainsi cités. Dans le but d'obtenir une représentation flexible de ce phénomène d'interaction, il serait judicieux de remplacer le vecteur poids par une fonction d'ensemble non additive, permettant ainsi de définir un poids non seulement sur chaque critère, mais aussi sur chaque sous-ensemble de critères (*e.g.,* $w_{\{Ab,Fr\}}$, $w_{\{Au,Ci\}}$). Dans l'exemple 1, comme les documents qui satisfont de manière équitable les deux critères sont les documents les plus préférés, on devrait attribuer un poids w_{12} au sous ensemble $\{c_1, c_2\}$. Idéalement, on doit assigner à w_{12} un score élevé (*e.g.*, $w_{12} = 1$) et attribuer à w_1 et w_2 des scores faibles (*e.g.*, $w_1 = w_2 = 0.35$) pour atténuer le score global dans le cas où les critères c_1 et c_2 ne répondent pas aux préférences (i.e., agissent de manière indépendante).

C'est dans ce but que l'utilisation des mesures floues a été proposée par (Sugeno, 1974) pour généraliser les mesures additives, étant donnée que dans de nombreuses situations du monde réel, l'additivité n'est pas une propriété appropriée pour les fonctions d'ensemble, à cause de son absence dans de nombreuses facettes du raisonnement humain (Marichal, 2000, 2002; Bouyssou *et al.*, 2006). Pour pouvoir exprimer la subjectivité humaine, Sugeno (1974) a proposé de remplacer la propriété d'additivité des fonctions d'ensemble par la monotonie à travers des mesures floues non additives (Grabisch, 1995).

C'est ainsi que nous proposons de traiter le problème d'agrégation de pertinence multidimensionnelle à l'aide de l'intégrale de Choquet (Choquet, 1953). Cette méthode d'agrégation est en outre caractérisée par sa capacité à généraliser plusieurs opérateurs d'agrégation classiques (Grabisch *et al.*, 2000) tels que la moyenne arithmétique pondérée ou les opérateurs OWA (Yager, 1988) et même les approches basées sur la combinaison linéaire (Vogt et Cottrell, 1999; Si et Callan, 2002). Cette généralisation sera discutée en plus de détails dans la section 5.3. En reprenant l'exemple déjà mentionné, les préférences souhaitées peuvent être simplement modélisées par une mesure floue μ tel que $\mu_{\{Ab,Fr\}} > \mu_{\{Au,Ci\}}$ où $\mu_{\{\cdot\}}$ représente le degré d'importance d'un critère (*resp.,* un sous ensemble de critères). Une des principales motivations de l'adaptation de cette famille des méthodes d'agrégation, est le fait que dans plusieurs travaux de RI, les dimensions de pertinence interagissent se sont avérées dépendants (Saracevic, 2007b; Carterette *et al.*, 2011). Par conséquent, l'intégrale de Choquet se basant sur le concept de mesure floue, s'avère un bon candidat pour ce type d'agrégation.

5.3 Cadre formel : l'opérateur de Choquet

L'intégrale de Choquet est une notion qui est apparue en 1953 (Choquet, 1953). Ses premières applications ont vu le jour dans les années 90 en aide multicritère à la décision (MCDA) (Murofushi et Soneda, 1993; Grabisch et Nicolas, 1994; Grabisch, 1995, 1996; Marichal, 1998). Depuis, cette intégrale non additive est devenue une branche importante de la théorie des fonctions d'agrégation et a été largement appliquée dans de nombreux domaines (Grabisch *et al.*, 2000). Dans cette section, nous présentons les concepts de base de cette méthode d'agrégation et nous donnons les principales caractéristiques qui pourront être utiles pour l'estimation de pertinence multidimensionnelle.

5.3.1 Concepts de base

Considérons l'ensemble $\mathcal{C} = \{c_1, \dots c_n\}$ des critères. Comme nous l'avons déjà montré, pour avoir une représentation flexible des phénomènes complexes d'interaction parmi ces critères (par exemple, une synergie positive ou négative entre certains critères), il faut pouvoir définir des poids non seulement sur chaque critère, mais aussi sur chaque sous-ensemble de critères.

Définition 3 Mesure floue**.**

Soit $I_{\mathcal{C}}$ l'ensemble de tous les sous ensembles de critère de $\mathcal{C}$. Une mesure floue est une fonction monotone μ de $I_{\mathcal{C}}$ à $[0, 1]$ tels que :
$\forall I_{C_1}, I_{C_2} \in I_{\mathcal{C}}$, si $(I_{C_1} \subseteq I_{C_2})$ alors $\mu(I_{C_1}) \leq \mu(I_{C_2})$, avec $\mu(I_{\varnothing}) = 0$ et $\mu(I_{\mathcal{C}}) = 1$.

Pour simplifier la notation, $\mu(I_{C_i})$ sera dénotée par μ_{C_i}. La valeur de μ_{C_i} peut être interprétée par le degré d'importance (ou le pouvoir de la *coalition*) de la combinaison de critères inclus dans le sous ensemble C_i. En effet, cette mesure généralise la notion de vecteurs de poids que nous avons présentée avec les méthodes classiques. Il est à noter que pour ces dernières (Cf., Section 5.2.2), le poids d'importance n'était défini que pour les critères seuls (singletons). Il suffisait donc de définir $n-1$ coefficients indépendants. Par contre, la mesure floue nécessite $2^n - 2$ coefficients pour être définie, devant vérifier les contraintes de monotonie. Dans le cas où la mesure est est k–additive, la mesure nécessite beaucoup moins de coefficients, *i.e.,* $\mu_A = 0$ pour tous les sous ensembles de critères $A \subseteq \mathcal{C}$ avec $|A| > k$. La figure 5.1 montre les différents valeurs de capacités à identifier dans le cas où le nombre de critères est égal à 3.

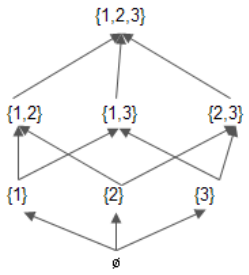


FIGURE 5.1: Exemple des différents valeurs de capacité à identifier dans le cas où le nombre de critères est égal à 3.

A partir de la mesure floue, nous pouvons donc construire une fonction d'agrégation floue permettant de calculer une sorte de valeur moyenne en prenant en compte les coefficients de la mesure floue (Sugeno, 1974, 1977). Nous discutons cet aspect dans la section suivante.

5.3.2 Principe d'agrégation et modélisation des interactions et corrélations

Nous présentons dans cette section le cadre général de l'agrégation multicritères avec l'intégrale de Choquet. Comme cette intégrale est considérée ici comme une fonction d'agrégation à N critères, nous adopterons la notation d'une simple fonction plutôt que la forme intégrale usuelle ($\int$), et l'intégrale sera un ensemble ordonné de N valeurs réelles obtenues sur un ensemble $\mathcal{A}$ d'alternatives. Le score global de $a_j \in \mathcal{A}$, donné par l'intégrale de Choquet selon une mesure floue μ et un ensemble $\mathcal{C}$ de critères, est défini par :

$$Ch_\mu(C_{1j}, \ldots, C_{Nj}) = \sum_{i=1,\ldots,N} c_{(i)j}(\mu_{C_{(i)}} - \mu_{C_{(i+1)}}) \tag{5.6}$$

avec $C_{(i)} = \{c_i, \ldots, c_N\}$ est un ensemble de critères avec $C_{(0)} = \varnothing$, $\mu_{C_{(0)}} = 0$ et $\mu_{C_{(1)}} = 1$ et $c_{(i)j}$ est le score obtenu selon un critère donné et $c_{(\cdot)j}$ indique que les indices sont permutés de façon à ce que $0 \leq c_{(1)j} \leq \ldots \leq c_{(N)j}$. Comme mentionné dans la section 5.2, C_{ij} est le score partiel [1] de a_j selon le critère c_i et $\mu_{C_{(i)}}$ est le degré d'importance de la combinaison $\{c_i, \ldots, c_N\}$ de critères.

Exemple 2 *Par exemple, si $C_{2j} \leq C_{1j} \leq C_{3j}$, nous aurons :*

$$\begin{aligned} Ch_\mu(C_{1j}, C_{2j}, C_{3j}) = & c_{(2)j}(\mu_{C_{2,1,3}} - \mu_{C_{(1,3)}}) \\ & + c_{(1)j}(\mu_{C_{1,3}} - \mu_{C_{(3)}}) \\ & + c_{(3)j}(\mu_{C_3}) \end{aligned}$$

La fonction d'agrégation de Choquet vérifie un certain nombre de propriétés naturelles et intéressantes : elle est continue, non décroissante, unanimement croissante, idempotente et interne. Une justification de toutes ces propriétés peuvent être trouvées dans (Grabisch, 1995; Grabisch *et al.*, 2000). En

1. La différence entre $c_{(i)j}$ et C_{ij} est que les scores partiels $c_{(i)j}$ sont permutés avant de calculer le score global, alors que C_{ij} est le score partiel de a_j, obtenu sur le critère c_i.

effet, de cette manière, l'intégrale de Choquet est bien capable de prendre en compte plusieurs types d'interactions entre les critères et représenter des préférences qui n'ont pas pu être capturées par un opérateur d'agrégation simple (Grabisch *et al.*, 2000). Notons que si μ est une mesure additive, l'intégrale de Choquet correspond à la moyenne pondérée. Des cas particuliers de l'opérateur de Choquet, dépendant de la mesure floue, incluent par exemple la moyenne arithmétique pondérée et l'opérateur OWA. Le tableau 5.1 présente quelque cas selon la valeur de la mesure de capacité.

	Intégrale de Choquet
OWA	$\mu_C = \sum_{j=0}^{i-1} w_{n-j}$, $\forall$ C tel que $\|C\| = i$, où $\|C\|$ dénote la cardinalité du sous ensemble de critères C.
MOYENNE ARITHMÉTIQUE PONDÉRÉE	Le poids w_i de chaque critère c_i est égal à (μ_{c_i}) et pour chaque sous ensemble de critères $C_1 \in \mathcal{C}$, $\mu_{C_1} = \sum_{c_i \epsilon C_1} \mu_{c_i}$
MOYENNE ARITHMÉTIQUE	$\mu_{C_1} = \frac{\|C_1\|}{\|\mathcal{C}\|}$

TABLE 5.1: Cas particuliers de l'intégrale de Choquet.

Pour faciliter la tâche d'interprétation sémantique du modèle résultat de l'intégrale de Choquet, nous introduisons deux paramètres intéressants appelés, *indice d'importance* et *indice d'interaction* (Murofushi et Soneda, 1993) qui permettent de traduire les relations ainsi que l'importance des critères. L'indice d'importance, appelé également indice de *Shapley* (Shapley, 1953), permet d'estimer la contribution moyenne qu'un critère (c_i) apporte à toutes les autres combinaisons de critères possibles. L'indice d'interaction permet de donner des informations sur le phénomène d'interaction pouvant exister entre un ensemble de critères. Nous définissons ces deux indices dans ce qui suit.

Définition 4 Indice de Shapley.

Soit μ_{c_i} le poids du critère c_i et $\mu_{Cr \cup c_i}$ sa contribution marginale à chaque sous ensemble de critères $Cr \in \mathcal{C}$. L'indice d'importance de c_i selon la mesure floue μ est défini comme la moyenne de toutes ces contributions :

$$\phi_\mu(c_i) = \sum_{Cr \subseteq \mathcal{C} \setminus \{c_i\}} \frac{(N-|Cr|-1)!.|Cr|!}{N!} [\mu_{Cr} \cdot \mu_{(Cr \bigcup c_i)}]$$

$\phi_\mu(c_i)$ mesure la contribution moyenne que (c_i) fournit à toutes les combinaisons de critères possibles.

L'indice d'importance ne donne aucune information sur le phénomène d'interaction pouvant exister entre les critères. L'importance globale de c_i ne peut pas être uniquement déterminés par son poids μ_{c_i}, mais aussi avec sa contribution marginale à tous les autres sous ensembles de critères. Alors, pour quantifier le degré d'interaction entre ces derniers, nous introduisons dans ce qui suit, le concept d'indice d'interaction.

Définition 5 Indice d'interaction.

Soit $(\Delta_{c_i c_j} \mu_{Cr})$, avec $Cr = \mathcal{C} \setminus \{c_i, c_j\}$, est la différence entre la contribution marginale du critère c_j à toute combinaison de critère contenant c_i, et une combinaison dans laquelle c_i est exclu.

$$(\Delta_{c_i c_j} \mu_{Cr}) = [\mu_{(\{c_i c_j\} \cup Cr)} - \mu_{(c_i \cup Cr)}] - [\mu_{(c_i \cup Cr)} - \mu_{Cr}]$$

Cette expression est définie pour estimer l'opposition entre deux critère c_i et c_j. Quand cette expression est positive (*resp.* négative) pour tout $Cr \in \mathcal{C} \setminus \{c_i, c_j\}$, on dit que les deux critères interagissent positivement (*resp.* négativement) (i.e., la contribution du critère c_j est plus significative avec la présence de c_i). L'interaction entre les deux mesures est alors définie comme suit :

$$I_\mu(c_i, c_j) = \sum_{Cr \subseteq \mathcal{C} \setminus \{c_i, c_j\}} \frac{(N-|Cr|-2)!.|Cr|!}{(N-1)!} (\Delta_{c_i c_j} \mu_{Cr})$$

Quand les deux critères sont indépendants, la valeur d'interaction, qui appartient à l'intervalle $[-1..1]$, est nulle. Dans le cas où les deux critères interagissent positivement (*resp.* négativement), la valeur est positive (*resp.* négative). Une fois que nous avons formalisé l'intégrale de Choquet et le concept de mesure floue, nous présentons dans la section 5.4 l'adaptation de cet opérateur pour l'estimation de pertinence multidimensionnelle.

5.4 iAggregator : un opérateur d'agrégation flou pour l'estimation de pertinence multidimensionnelle

5.4.1 Modèle d'agrégation

Nous introduisons dans cette section notre approche basée sur l'intégrale de Choquet pour l'agrégation de pertinence multidimensionnelle. La combinaison des critères de pertinence revient à agréger les scores partiels (*RSV* pour *Retrieval Status Value*) issus de chaque critère. Le choix de l'opérateur de Choquet pour ce type de problème est dû principalement aux spécificités suivantes :

- En RI multicritères, les critères de pertinence peuvent être dépendants ou corrélés, comme déjà montré dans de nombreux travaux de l'état de l'art (Saracevic, 2007a; Carterette *et al.*, 2011; Eickhoff *et al.*, 2013b) ;
- La possibilité de définir des importances relatives entre les critères (Grabisch *et al.*, 2000) ;
- La possibilité d'exprimer des interactions entre les critères, comme les effets de redondance ou de synergies ;
- La facilité d'interprétation sémantique des valeurs liées à l'importance accordée au critères, ceci permet d'éviter l'effet *boite noire*, qui existe, par exemple, lors de l'application des algorithmes d'apprentissage d'ordonnancement (Liu, 2009).

Nous présentons dans la suite, notre modèle d'agrégation, appelé iAggregator [2], pour l'estimation de pertinence multidimensionnelle. Soient $\mathcal{D}$ la collection des documents, le score global de $d_j \in \mathcal{D}$, donné par notre modèle selon une mesure floue μ et un ensemble $\mathcal{C}$ de critères de pertinence, en réponse à une requête q, est défini par :

$$RSV_{\mathcal{C}}(q, d_j) = Ch_{\mu}(RSV_{c_1}(q, d_j), \ldots, RSV_{c_N}(q, d_j)) \tag{5.7}$$

$$= \sum_{i=1}^{N} rsv_{(i)j} \cdot (\mu_{\{c_i,\ldots,c_N\}} - \mu_{\{c_{i+1},\ldots,c_N\}}) \tag{5.8}$$

2. iAggregator : pour *interactive Aggregator*, afin de souligner la propriété d'interaction qui distingue cet opérateur d'agrégation des opérateurs classiques.

avec $rsv_{(i)j}$ est le score obtenu selon un critère donné [3], $C_{(i)} = \{c_i, ..., c_N\}$ est un ensemble de dimensions de critère de pertinence, avec $C_{(0)} = \varnothing$, $\mu_{C_{(0)}} = 0$ et $\mu_{C_{(1)}} = 1$.

D'un point de vue théorique, l'intégrale de Choquet dispose d'un nombre de propriétés qui semblent être pertinentes pour un domaine tel que la RI ; étant donné qu'elle est construite à partir du concept de mesure floue, elle permet la modélisation des relations d'interaction flexibles en permettant la modélisation des relations de dépendance complexes entre les critères (Grabisch *et al.*, 2000). Nous distinguons trois types d'interactions, illustrées par la figure 5.2. Les axes d'abscisses et d'ordonnées représentent les scores d'évaluations des quatre documents d_1, d_2, d_3 et d_4 suivant les deux critères c_1 et c_2, respectivement. Les documents liés par des lignes pointillés ont le même degré de pertinence.

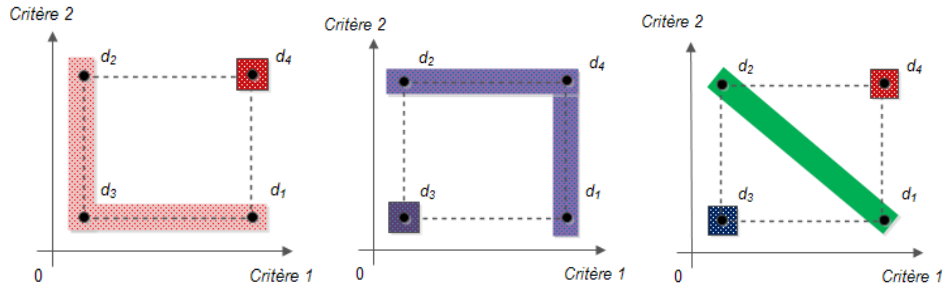


FIGURE 5.2: Interactions possibles entre les critères de pertinence.

– *Interaction positive*, appelée aussi *synergie positive*, quand le poids global de deux critères est supérieure à leur poids individuels : $\mu_{\{c_i,c_j\}} > \mu_{c_i} + \mu_{c_j}$. Cette inégalité peut être interprétée comme suit : *"la contribution de c_j à toute combinaison de critères contenant c_i est strictement supérieure à la contribution de c_j à la même combinaison quand c_i est exclu"*. Dans ce cas, c_i et c_j sont négativement corrélés, i.e., la satisfaction d'un critère unique doit produire un impact très faible par rapport à la satisfaction des deux critères ensemble. Intuitivement, dans un contexte de RI, cette propriété favorise les documents qui sont satisfaits équitablement par tous les ensembles de critères, plutôt que les documents sur-estimés selon un seul critère de pertinence. Dans ce cas, les critères peuvent être également

3. $rsv_{(\cdot)j}$ indique que les indices sont permutés de façon à ce que $0 \leq rsv_{(1)j} \leq \ldots \leq rsv_{(N)j}$.

considérés comme présentant un degré de complémentarité ou d'opposition. Par exemple, dans la figure 5.2a, le document d_4 doit être préféré aux documents d_2 et d_3, étant donnée qu'ils satisfont pas équitablement les deux critères c_1 et c_2.

– *Interaction négative* (*synergie négative*), quand le poids global de deux critères est plus petit que leurs poids individuels : $\mu_{\{c_i,c_j\}} < \mu_{c_i} + \mu_{c_j}$. Dans ce cas, on peut dire que l'union des critères n'a pas de valeur ajouté sur l'évaluation globale des documents, i.e., la contribution marginale de c_j à chaque combinaison de critères contenant c_i est strictement inférieure à la contribution marginale de c_j à cette même combinaison mais où c_i est exclu. Ces deux critères présentent alors une sorte de redondance. Cette spécificité est parmi les points clés de l'intégrale de Choquet, vu qu'elle permet d'absorber le biais qui pourrait être introduit par l'implication des critères de pertinence redondants dans l'évaluation globale des documents. Ceci est effectué par l'association d'un degré d'importance μ_{c_i,c_j} relativement faible au sous ensemble des critères positivement corrélés. Nous remarquons à partir de la figure 5.2b que le document d_4 a la même importance que les documents d_2 et d_3, vu que la satisfaction des critères c_1 ou c_2, qui sont redondants, est suffisante pour juger un document comme pertinent.
– *Indépendance*, quand il n'existe aucune corrélation entre l'ensemble des critères. Dans ce cas, on dit que la mesure floue est additive : $\mu_{\{c_i,c_j\}} = \mu_{c_i} + \mu_{c_j}$. La moyenne arithmétique pondérée est exemple de ce type de fonction qui permet l'indépendance des critères. Le poids de chaque critère indique son importance relative.

Après avoir présenté les méthodes permettant l'interprétation et la compréhension du modèle résultant de l'intégrale de Choquet, nous introduisons une des questions les plus complexes dans les problème d'agrégation avec l'intégrale de Choquet, et les mécanismes d'agrégation supervisés en général. Ce défi consiste en l'identification des mesures floues (ou encore les poids d'importance des critères), surtout quand le nombre de critères est élevé (Grabisch *et al.*, 2008). En effet, la complexité ici n'est pas liée au processus d'agrégation en soi, mais plutôt à l'identification des valeur de capacité, qui est une étape d'apprentissage bien avant l'application de l'opérateur, contrairement aux algorithmes d'apprentissage d'ordonnancement où la complexité est à la fois dans l'apprentissage que l'ordonnancement. Pour répondre à ce défi, nous allons nous baser sur la méthode des moindres carrés ; une des mé-

thodes d'optimisation les plus utilisées dans la littérature [4] (Grabisch *et al.*, 2008). La méthode d'apprentissage est détaillé dans la section suivante.

5.4.2 Apprentissage des poids d'importances

Notation	Description
Q_{app}	L'ensemble des requêtes utilisées pour apprendre les valeurs de capacités.
N	Nombre de critères de pertinence.
$\mathcal{D}$	La collection de documents.
K	Nombre de documents utilisés pour l'apprentissage pour chaque requête.
$\gamma^{i,r}$	Liste ordonnée de documents en réponse à la requête q_r suivant la combinaison de capacité $\mu^{(i)}$. Soit $P@X(\gamma^{r,i})$ la $P@X$ de $\gamma^{r,i}$ et $AVP@X(\gamma^i)$ soit sa moyenne de $P@X$ sur toutes les requêtes $\in Q_{app}$ suivant $\mu^{(i)}$.
I_{Cr}	Tous les sous ensembles de critères possibles de Cr.
$\mathcal{S}_\mu$	Ensemble de combinaisons de capacité expérimentées. Chaque combinaison $\mu^{(i)} \in \mathcal{S}_\mu$ contient les valeurs de capacités de tous les ensembles et sous ensemble de critères.

TABLE 5.2: Synthèse des notations utilisées avec l'algorithme 1.

L'objectif de la phase d'apprentissage est de paramétrer les mesures floues selon une mesure objective de RI (e.g. $P@X$) en identifiant les valeurs de capacité. Nous proposons dans ce qui suit un algorithme générique permettant d'apprendre ces capacités indépendamment du nombre de critères de pertinence, et de la tâche de RI considérée.

Les données d'apprentissage nécessaires pour identifier les mesures floues de l'intégrale de Choquet comprennent un ensemble de requêtes d'apprentissage, et pour chaque requête, un ensemble ordonné de documents représentés par des vecteurs contenant des scores partiels selon chaque critère ; chaque document est annoté avec une étiquette (*e.g.,* pertinent ou non pertinent). La méthodologie adoptée est détaillée dans l'algorithme 1. Le Tableau 6.1 décrit les notations utilisées dans cet algorithme. Ce dernier comprend deux étapes principales :

4. Nous avons utilisé cette méthode au sein du package KAPPALAB sous R (Grabisch *et al.*, 2008).

Algorithm 1: Apprentissage des mesures floues

Entrées: Q_{learn}, N, K.
Sortie: Combinaison de capacité optimale $\mu^{(**)}$.

Étape 1 : Initialisation des valeurs de capacités

$m \leftarrow (2^N - 1) \times N$;

1. **Pour** $i = 1$ à m {*Identification des combinaisons de capacités*} **Faire**
2. $\mu^{(i)} = (\bigcup_{j:1..N} \{\mu_{c_j}\}) \bigcup (\bigcup_{Cr \in \mathcal{C}, |Cr|>1} \{\mu_{I_{Cr}}\})$; $\mu_{I_{Cr}} = \sum_{c_i \in Cr, |c_i=1|} \mu_{c_i}$
3. **Fin Pour**
4. **Si** $N \geq 4$ {*Supposer la 2-additivité*} **Alors**
5. **Pour** chaque $I_{Cr} \in \mu^{(i)}$ tel que $|Cr| > 2$ **Faire**
6. $\mu_{I_{Cr}} = 0$
7. **Fin Pour**
8. **Fin Si**
9. $\mathcal{S}_\mu = \bigcup_{i:1..m} \{\mu^{(i)}\}$
10. **Pour** chaque $\mu^{(i)} \in \mathcal{S}_\mu$ {*paramétrage des capacités*} **Faire**
11. Calculer $AVP@X(\gamma^i)$
12. **Fin Pour**
13. $Cmax = \underset{1...|\mathcal{S}_\mu|}{\text{Argmax}} (AVP@X(\gamma^i))$; $\mu^{(*)} = \mu^{(cmax)}$

Étape 2 : Optimiser les valeurs de capacités

14. $D^{app} = \varnothing$
15. **Pour** $r = 1$ à $|Q^{app}|$ {*Interpoler les scores globaux*} **Faire**
16. $D^{app} = D^{app} \cup \gamma^{*,r}$
17. **Pour** $j = 1$ à K **Faire**
18. $RSV^{int}_{\mathcal{C}}(q_r, d_j) = \underset{1...d'_j \in \gamma^{*,r}, d'_j >_{\mathcal{C}} d_j}{\text{Max}} (RSV_{\mathcal{C}}(q_r, d'_j))$; $\gamma^{*,r} = \gamma^{*,r} \setminus \{d_j\}$
19. **Fin Pour**
20. **Fin Pour**

{*Optimisation basée sur la méthode des moindres carrés.*}

21. **Répéter**
22. $\mathcal{F}_{LS}(\mu) = \sum_{d_j \in D^{app}} [Ch_\mu(\text{RSV}_{c_1}(d_j), \ldots, \text{RSV}_{c_N}(d_j)) - \text{RSV}^{int}_{\mathcal{C}}(d_j)]^2$
23. **Jusqu'à** convergence
24. **Retourner** le résultat $\mu^{(**)}$

– *Initialisation des valeurs des combinaisons de capacités.* Une combinaison de capacités $\mu^{(.)}$ désigne un ensemble des valeurs de capacités associées à chaque critère et à chaque sous-ensemble de critères. Par exemple, dans le

cas de trois critères de pertinence, une combinaison de capacités comprend ($\{\mu_{c_1}; \mu_{c_2}; \mu_{c_3}; \mu_{c_1,c_2}; \mu_{c_1,c_3}\ ;\ \mu_{c_2,c_3}\}$). Afin de paramétrer ces valeurs, nous utilisons une mesure de RI telle que la $P@X$ sur les requêtes d'apprentissage Q^{app}. Le paramétrage est concevable en RI multicritères, étant donné que le nombre de critères de pertinence est généralement petit (Saracevic, 2007b). Cependant, lorsque le nombre de critères est supérieur ou égale à 4, la méthode devient plus complexe, mais nous pouvons éviter la complexité du paramétrage en se basant sur la famille des capacités 2-additive (Grabisch *et al.*, 2000) nécessitant moins de coefficients à définir.

– *Optimisation des valeurs de capacités.* En partant d'une combinaison de capacités $\mu^{(*)}$ obtenue dans l'étape précédente, on extrait les K premiers documents retournés en réponse à chaque requête $q \in Q^{app}$. Les scores de ces documents (D^{app}) sont interpolés afin de placer les documents non pertinents à la fin de la liste l'ordonnancement. Après avoir obtenu les scores de pertinence globaux désirés $\text{RSV}_{\mathcal{C}}^{int}(q, d_j)$ pour chaque document $d_j \in D^{app}$, et étant donné que nous disposons des étiquettes $\text{RSV}_{c_i}(q, d_j)$, nous procédons à l'application de la méthode des moindres carrés pour l'identification des valeurs de capacités des critères et des sous-ensembles de critères considérés.

5.5 Personnalisation de la méthode d'agrégation de pertinence

Dans la section 5.4, nous avons exploité l'opérateur de Choquer pour quantifier explicitement l'importance absolue des dimensions de pertinence et pondérer et agréger ainsi les scores de différentes dimensions de pertinence. Dans cette partie, nous tournons le problème d'agrégation pour tenir compte de la propriété de subjectivité qui peut se décliner à travers les différences entre les préférences utilisateurs quant à l'importance accordée à chaque dimension de pertinence. Le défi dans le problème d'agrégation dans le cadre de RI personnalisée est :

1. L'estimation de l'importance des critères : identifier les critères devant avoir un poids d'importance plus élevé que d'autres ;
2. L'agrégation : combiner efficacement les critères de pertinence en tenant compte des dépendances pouvant exister entre eux ;
3. Apprendre les mesures d'importance des critères pour chaque utilisateur ;

La fonction d'agrégation de pertinence personnalisée basée sur l'intégrale de Choquet est définie comme suit :

Définition 6 Agrégation personnalisée.

$RSV^u_{\mathcal{C}}(q,d_j)$ est le score de pertinence personnalisé de d_j pour l'utilisateur u suivant l'ensemble des critères de pertinence $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ défini comme :

$$RSV^u_{\mathcal{C}}(q,d_j) = Ch_\mu(RSV^u_{c_1}(q,d_j), \ldots, RSV^u_{c_N}(q,d_j)) = \sum_{i=1}^{N} \mu^u_{\{c_i,\ldots,c_N\}} \cdot (rsv^u_{(i)j} - rsv^u_{(i-1)j})$$

Où Ch_μ la fonction d'agrégation de Choquet, $rsv^u_{(i)j}$ est le $i^{\text{ème}}$ élément de la permutation $RSV(q,d_j)$ sur le critère c_i, tel que $(0 \leq rsv^u_{(1)j} \leq ... \leq rsv^u_{(N)j})$, $\mu^u_{\{c_i,\ldots,c_N\}}$ est le degré d'importance de l'ensemble des critères $\{c_i, ..., c_N\}$ pour l'utilisateur u.
De cette manière, nous sommes capables d'ajuster les paramètres du modèle d'ordonnancement automatiquement pour chaque utilisateur, rendant ainsi les résultats dépendant de ses préférences sur les critères considérés.

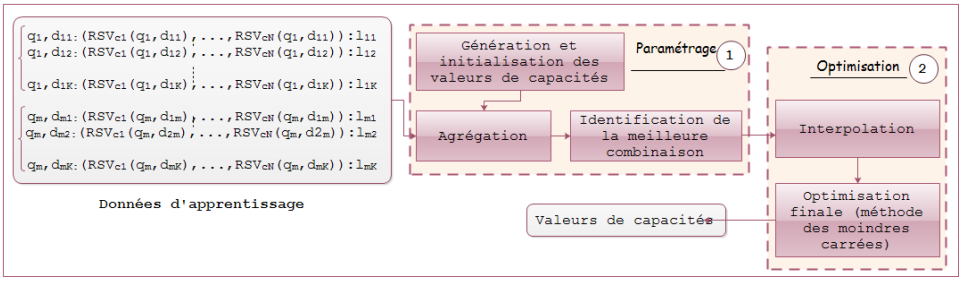


FIGURE 5.3: Les différentes étapes pour l'apprentissage des valeurs de capacités.

L'apprentissage des poids d'importance des critères pour chaque utilisateur est appliqué avec le même algorithme d'identification des capacités présenté dans la section 5.4.2. Ceci permet également d'ajuster les degrés d'importance de chaque critère de pertinence selon les préférences de l'utilisateur en question. En revanche, dans le cas d'un contexte de RI non personnalisée, cet algorithme est appliqué indifféremment pour toutes les requêtes issues de la tâche, permettant de quantifier des scores de pertinences génériques liées à la tâche et non spécifiquement aux utilisateurs. On calcule pratiquement, comme déjà montré dans la section précédente, $\text{RSV}c_i(q,d_j)$ au lieu de $\text{RSV}^u_{c_i}(q,d_j)$. La figure 5.3 montre les différentes étape appliquées pour

apprendre les mesures d'importance pour chaque utilisateur.

5.6 Évaluation expérimentale

5.6.1 Objectifs

Les objectifs de ces expérimentations sont :

1. Montrer les corrélations existantes entre les critères considérés par le biais de la mesure floue ;
2. Évaluer l'efficacité de l'intégrale de Choquet comparativement à d'autres opérateurs d'agrégation standards dans des cadres et des collections de RI différentes.

5.6.2 Cadres d'évaluation

5.6.2.1 Tâche 1 : recherche de tweets

Dans cette section, nous évaluons notre approche d'agrégation de pertinence dans un cadre de RI sociale. Plus particulièrement, nous considérons une tâche de recherche de tweets, dans le cadre de la tâche Microblog de TREC, où la pertinence des documents est basée sur trois dimensions de pertinence à savoir, la dimension thématique, la fraîcheur des tweets, et l'autorité des auteurs des tweets (Duan *et al.*, 2010; Nagmoti *et al.*, 2010).

5.6.2.1.1 Description de la tâche de recherche Comme défini dans la tâche Microblog de TREC 2011 (Ounis *et al.*, 2011), la recherche de tweets est une tâche en temps réel dans laquelle les utilisateurs s'intéressent à l'information pertinente et récente, à la fois. Des travaux récents s'intéressant à cette tâche de recherche ont identifié plusieurs facteurs ayant un impact considérable dans le classement final des documents (Nagmoti *et al.*, 2010). Parmi ces critères nous citons : la *topicalité*, la longueur des tweets, la présence des *URLs* dans un tweet et l'autorité (Duan *et al.*, 2010; Chen *et al.*, 2012). En effet, la spécificité de la tâche Microblog de TREC 2011 (Ounis *et al.*, 2011) qui définit la recherche de tweets comme étant une tâche de recherche dans laquelle les utilisateurs s'intéressent aux informations récentes

et pertinentes, motive bien l'utilisation de l'intégrale de Choquet, où les interactions entre les critères est bien prise en considération. Ceci est d'autant plus important que les utilisateurs préféreront les tweets qui sont à la fois pertinents et récents, *i.e.,* dont les scores sont balancés entre ces deux critères et non pas biaisés par l'un des deux. Pour notre part, nous exploitons trois dimensions de pertinence à savoir, la topicalité (thématique de recherche) (To), la fraîcheur des tweets (Fr), et l'autorité (Au). Nous signalons ici que notre objectif dans ce travail n'est pas de chercher à identifier tous les critères *utiles* pour la recherche de tweets. Il existe de nombreux travaux qui se sont déjà intéressés à l'exploitation des facteurs de pertinence dans les *microblogs* Nagmoti *et al.* (2010); Duan *et al.* (2010); Carterette *et al.* (2011); Berardi *et al.* (2011a); Metzler et Cai (2011); Cheng *et al.* (2013); Amati *et al.* (2012); Choi et Croft (2012); Damak *et al.* (2013); Miyanishi *et al.* (2014). L'objectif principal de notre étude ici est d'optimiser la fonction d'agrégation de pertinence, indépendamment de la tâche de RI et des critères exploités. Nous avons cependant utilisé les critères les plus utilisés pour cette tâche de recherche de tweets. L'agrégation de ces critères avec l'intégrale de Choquet selon une mesure floue μ, en réponse à une requête utilisateur Q, est définie par :

$$Ch_\mu(To(T_j,Q), Au(T_j), Fr(T_j)) = \sum_{i=1,...,3} (c_{(i)j} - c_{(i-1)j}).\mu_{C_{(i)}} \tag{5.9}$$

où T_j est un *tweet*, $c_{(i)}$ indique qu'un score partiel [5] obtenu sur un critère donné, en réponse à la requête Q, sont permutés tel que : $0 \leq c_{(1)j} \leq c_{(2)j} \leq c_{(3)j}$, et $C_{(i)} = \{c_i, ..., c_3\}$. Ch_μ est le score global qui définit le classement final de chaque document selon les trois dimensions de pertinence. Nous donnons dans ce qui suit la formalisation de ces différents critères.

– *Topicalité :* nous proposons d'utiliser la fonction Okapi BM25 (Robertson et Jones, 1976) pour le classement des tweets en réponse à une requête Q donnée :

$$To(T,Q) = BM25(T,Q) = \sum_{q_i \in Q} \frac{Idf(q_i) \cdot tf(q_i,T) \cdot (k_1+1)}{tf(q_i,T) + k_1(1 - b + b\frac{Length(T)}{avg_{length}})} \tag{5.10}$$

où q_i est un terme de la requête, $Idf(q_i)$ est la fréquence inverse du document, $Length(T)$ dénote la longueur d'un *tweet* T et avg_{length} représente la moyenne des longueurs des tweets dans la collection.

– *Autorité :* représente l'influence des auteurs des tweets dans Twitter. Nous la définissons comme présentée dans (Nagmoti *et al.*, 2010) :

5. Tous les scores partiels sont normalisés et sont tous dans $[0, 1]$.

$Au(T) = Au_{nb}(T) + Au_{me}(T)$, où *(i)* $Au_{nb}(T)$ est le *nombre total des tweets*, pour favoriser les tweets publiés par des utilisateurs influents. Nous la définissons par : $Au_{nb}(T) = N(a_i(T))$, avec $a_i(T)$ représentant l'auteur du *tweet* T, et $N(a_i(T))$ dénotant le nombre de tweets publiés par a_i. *(ii)* $Au_{me}(T)$ est le *nombre de mentions d'un auteur*, *i.e.,* plus un auteur est mentionné, plus il est populaire. Elle est définie par : $Au_{me}(T) = N(ma_i)(T))$, avec $N(ma_i)(T))$ estimant le nombre de fois l'auteur du *tweet* T est mentionné.

– *Fraîcheur des tweets :* c'est la différence entre le temps de publication du *tweet* $Tp(T)$ et la date de soumission de la requête $Ts(Q)$: $Fr(T) = T_s(Q) - T_p(T)$.

5.6.2.1.2 Données expérimentales Nous exploitons ici la collection de tweets fournie par la tâche Microblog de TREC 2011 et TREC 2011 (Ounis *et al.*, 2011, 2012). La collection inclut environ 16 millions de tweets publiés sur 16 jours. Les statistiques sont données dans le Tableau 5.3.

Tweets	16,141,812
Tweets null	1,204,053
Termes uniques	7,781,775
Nombre de twitters	5,356,432
Nombre de Topics de TREC Microblog 2011	49
Nombre de Topics de TREC Microblog 2012	60

TABLE 5.3: Statistiques de la collection fournie par la tâche Microblog de TREC 2011 et 2012.

La figure 5.4 présente un exemple de topic de la tâche Microblog de TREC 2012.

```
<top>
<num> Number: MB106 </num>
<query> Steve Jobs' health </query>
<querytime> Tue Feb 08 10:05:10 +0000 2011 </querytime>
<querytweettime> 34915635595059201 </querytweettime>
</top>
```

FIGURE 5.4: Exemple de topic de la tâche Microblog de TREC 2012.

5.6.2.2 Tâche 2 : recherche de lieux d'attractions

5.6.2.2.1 Description de la tâche de recherche La tâche "Contextual Suggestion" de TREC a pour objectif d'évaluer les techniques de recherche répondant à des besoins en information, qui sont fortement tributaires du contexte dans lequel se situe un utilisateur. Étant donné un utilisateur, cette tâche a pour objectif de chercher des places d'attractions (eg., restaurants, parcs d'attractions, zoo, etc.) pouvant l'intéresser suivant deux critères de pertinence : *(i)* les centres d'intérêt de l'utilisateur, *i.e.,* ses préférences sur un historique de recherche de places ; *(ii)* sa localisation géographique.

5.6.2.2.2 Données expérimentales La collection de test présente les caractéristiques suivantes :

– *Utilisateurs :* le nombre total d'utilisateurs est égal à 635. Chaque utilisateur est représenté par un profil reflétant ses préférences sur des lieux d'une liste de 50 exemples de suggestions. Un exemple de suggestion est un lieu d'attraction qui est susceptible d'intéresser l'utilisateur. Chaque exemple est représenté par le titre du lieu, une brève description et une *URL* du site web correspondant. La figure 5.5 présente un exemple de suggestion.

```
<ID> 65 </ID>
<Title> Red Mango </Title>
<Description> Red Mango is committed to providing the
healthiest and best tasting all-natural nonfat frozen
yogurt and fresh fruit smoothies. No wonder Zagat ranked us
#1, twice. </Description>
<URL> http://www.redmangousa.com </URL>
```

FIGURE 5.5: Exemple de suggestion de lieu d'attraction.

Les préférences des utilisateurs sont données sur une échelle de 5 points et sont attribuées aux descriptions et aux *URLs* des exemples de suggestions. Les préférences positives (*resp.,* négatives) sont celles ayant un degré de pertinence égal à 3 ou à 4 (*resp.,* 0 ou 1) selon la description du site et la correspondance par rapport à l'*URL*. Un exemple de préférence sur l'exemple de suggestion 65 pour l'utilisateur 534 pourrait être : $534, 65, 4, 4$ avec 4 et 4 sont respectivement, les préférences sur l'*URL* `http://www.redmangousa.com` et la description donnée avec.

– *Contextes :* le nombre de contextes fournis est égal à 50 ; chaque contexte correspond à une position géographique dans une ville donnée. La position géographique est décrite par une longitude et une latitude. Étant donnés une paire d'utilisateur et un contexte représentant la requête, l'objectif principal de la tâche est de fournir une liste de 50 suggestions triées par ordre de pertinence selon les critères centres d'intérêt et géolocalisation. Un exemple de contexte relatif à la ville *Monroe, LA* comprend : 71, *Monroe, LA*, 32.81513 , –92.20569 où les deux derniers attributs correspondent à la position géographique de l'utilisateur.
– *Collection de documents :* pour chercher des suggestions de lieux à partir du web, nous avons exploité l'API Google Place [6]. Comme pour la plupart des groupes participants à la tâche "Contextual Suggestion" (Dean-Hall *et al.*, 2013), nous commençons par interroger l'API Google Place avec les requêtes appropriées en se basant sur la localisation géographique des lieux. Étant donné que l'API Google Place renvoie jusqu'à 60 suggestions par requête, nous avons effectué une nouvelle recherche avec des paramètres différents tels que les types de lieux qui sont pertinents par rapport à la tâche (*e.g.,* restaurant, pizzeria, musée, etc.). Nous avons collecté, en moyenne, environ 157 suggestions par requête et 3925 suggestions au total. Pour obtenir les scores des documents collectés selon le critère de géolocalisation, nous avons calculé la distance entre les lieux collectés et le contexte. Les scores des documents selon le critère centres d'intérêts sont calculés en se basant sur le cosinus de similarité entre la description des suggestions et le profil de l'utilisateur. Les profils des utilisateurs sont représentés par des vecteurs de termes construits à partir de leurs préférences personnelles sur les exemples de suggestions. La description des lieux est construite à partir des *snippets* des résultats renvoyés par le moteur de recherche Google [7] lorsque l'URL du lieu est soumise sous forme d'une requête.
– *Jugements de pertinence :* les jugements de pertinence de cette tâche sont effectués par les utilisateurs et mandatés par TREC à la fois (Dean-Hall *et al.*, 2013). Chaque utilisateur représenté par un profil, juge les lieux qui lui sont suggérés de la même façon que les exemples de suggestions. Ainsi, l'utilisateur affecte un jugement de $0-4$ à chaque titre/description et à chaque *URL*, tandis que les assesseurs de TREC jugent les suggestions uniquement en termes de correspondance au critère géolocalisation avec une évaluation de (2, 1 et 0). Une suggestion est considérée comme

6. `https://developers.google.com/places`
7. `https://www.google.com`

pertinente si elle a un degré de pertinence égal à 3 ou 4 selon le critère centre d'intérêts (profil) et une évaluation égale à 1 ou 2 selon le critère géolocalisation. Dans ce qui suit, ces jugements de pertinence constituent notre réalité de terrain utilisée pour l'apprentissage et le test.

5.6.2.3 Tâche 3 : recherche dans les folksonomies

Dans cette section, nous décrivons le cadre d'évaluation proposé dans le contexte d'une RI dans les folksonomies. Nous avons exploité une collection de 33k signets [8] collectés à partir du système d'annotation *Del.icio.us* [9]. Le corpus inclut des informations d'évaluation données pour 35 utilisateurs en réponse à 177 requêtes (Vallet et Castells, 2012), suivant deux critères : thématique (To) des signets étant données une requête et leur pertinence personnelle (Us) pour un utilisateur spécifique.

5.6.3 Évaluation de iAggregator dans un contexte de RI sociale

5.6.3.1 Protocole d'évaluation

Dans l'évaluation de notre modèle dans la tâche de recherche de tweets, nous avons exploité 49 requêtes de la tâche Microblog de TREC 2011 pour l'apprentissage des mesures floues et nous avons utilisé les 60 requêtes de TREC 2012 pour le test. Trois critères de pertinence liés à la tâche ont été utilisés pour le calcul des scores des documents (Nagmoti *et al.*, 2010) : sujet des requêtes, fraîcheur de l'information et autorité du *twitterer*.

Pour évaluer les performances de notre approche dans ce cadre, nous avons comparé les résultats issus de l'application de l'opérateur proposé aux méthodes les plus utilisées en RI pour la combinaison de pertinence multicritères :

– Moyenne arithmétique (Am) et moyenne arithmétique pondérée (Wam) ;
– Méthode de combinaison linéaire classique (Mcl) (Vogt et Cottrell, 1999; Larkey *et al.*, 2000; Si et Callan, 2002; Craswell *et al.*, 2005) ;
– Opérateurs Min et Max ;
– Opérateurs Owa (Yager, 1988) et OWmin (Dubois et Prade, 1996) ;

8. `http://ir.ii.uam.es/~dvallet/persdivers/index.htm`
9. `http://www.delicious.com`

– Opérateurs prioritaires AND et SCORING (da Costa Pereira *et al.*, 2009) ;
– Algorithmes d'apprentissage d'ordonnancements : RANKNET (Burges *et al.*, 2005), RANKSVM (Joachims, 2006), LISTNET (Cao *et al.*, 2007), RANDOM FOREST (Breiman, 2001) et λ–MART (Wu *et al.*, 2010).

Les mesures d'évaluation utilisées sont la précision (P@5, P@10, P@20, P@30) et la précision moyenne (MAP). Nous avons calculé ces mesures avec l'outil standard *trec_eval*[10]. La mesure P@30 est la mesure officielle des tâches TREC Microblog 2011 et 2012.

5.6.3.2 Paramétrage et identification des mesures floues

Comme nous disposons des jugements de pertinence (*qrels*) associés aux différentes requêtes de la tâche, nous nous sommes basés sur la méthode détaillée dans l'algorithme 1 pour paramétrer les mesures floues associées aux trois critères utilisés dans cette tâche. En effet, nous avons tout d'abord testé un ensemble de combinaisons de capacités, comme précisé dans l'étape 3 de l'algorithme 1, tel que la somme des trois scores individuels partiels soit égale à 1. Dans le cas où nous avons trois critères de pertinence, chaque combinaison de capacités est composée d'un ensemble de valeurs qu'on peut assigner à tous les sous ensembles de critères possibles :

$$\mu^{(i)}(c_{\text{topicalité}}, c_{\text{autorité}}, c_{\text{fraîcheur}}) = \{\mu_{\{\text{topicalité}\}}, \mu_{\{\text{autorité}\}}, \mu_{\{\text{fraîcheur}\}}, \mu_{\{\text{topicalité,autorité}\}}, \mu_{\{\text{topicalité,fraîcheur}\}}, \mu_{\{\text{fraîcheur,autorité}\}}\}$$

Les valeurs des capacités des critères sont paramétrées de façon à ce que la somme des valeurs partiels soit égale à 1. Ainsi, nous avons testé 21 combinaisons de valeurs possibles obtenues comme suit :

1. Nous avons commencé par attribuer la plus grande valeur (i.e., 0.8) au critère topicalité. Les valeurs des capacités des critères fraîcheur et autorité sont égales à 0.1 (i.e., la somme des valeurs des trois critères est égale à 1). Les valeurs des capacités des sous ensembles de critères sont égales à la somme de leurs valeurs individuelles. Ensuite, nous avons décrémenté la valeur de capacité du critère topicalité de 0.1 et nous avons respectivement incrémenté de 0.1 celles de la fraîcheur et de l'autorité. Ce processus est répété jusqu'à ce que la valeur de capacité de la fraîcheur atteigne 0.8.

10. `http://trec.nist.gov/trec_eval`

2. Nous associons une valeur de capacité plus élevée au critère fraîcheur (i.e., 0.8), et nous incrémentons celle de l'autorité de 0.1, jusqu'à ce qu'elle atteigne 0.8.
3. Nous décrémentons la valeur de la capacité d'autorité et nous incrémentons celle du critère topicalité jusqu'à ce qu'elle atteigne 0.8.

Pour ce faire, nous nous sommes basés sur l'ensemble des requêtes de la tâche Microblog 2011 pour l'apprentissage. Nous notons cet ensemble par Q_{app}. La figure 5.6 présente des exemples de capacités avec les résultats obtenus en terme de $P@30$ pour chacune des combinaisons, et ce après application de l'intégrale de Choquet sur les trois dimensions de pertinence (To, Au, Fr). La combinaison de capacités $\mu^{(1)}$ donnée dans la figure 5.6, représente la combinaison optimale obtenue dans la phase d'apprentissage donnant ainsi la meilleure valeur de $P@30$ sur l'ensemble Q_{app}.

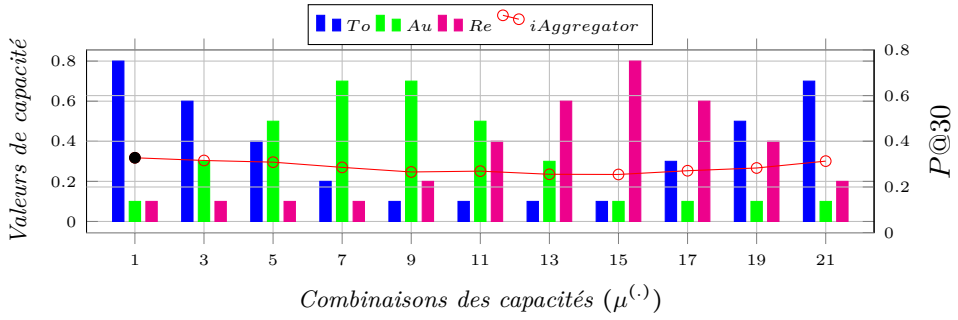


FIGURE 5.6: Paramétrage des valeurs de capacités dans la tâche de recherche de tweets et résultats des valeurs de précisions pour les combinaisons de capacités utilisées pour l'apprentissage. L'axe des abscisses représente quelque combinaisons parmi les 21 utilisées. L'axe des ordonnées à droite présente les valeurs de précision, et l'axe à gauche présente les valeurs de capacités pour les critères de pertinence.

La combinaison $\mu^{(1)}$ inclue les valeurs suivantes : (μ_{To} = 0.8, μ_{Au} = 0.1, μ_{Re} = 0.1, $\mu_{To,Au}$ = 0.9, $\mu_{To,Re}$ = 0.9, $\mu_{Au,Re}$ = 0.2). Comme nous pouvons le voir à partir des valeurs de $\mu^{(1)}$ et des performances données par les autres capacités dans la figure 5.6, notre modèle est pénalisé pour les préférences sur les tweets, pour lesquels uniquement les critères thématique de recherche et autorité se voient attribuer des valeurs de capacités importantes. Néanmoins, la précision est meilleure pour les tweets pour lesquels le critère fraîcheur d'information et topicalité sont importants, ce qui implique peut être que

les requêtes traitent des sujets dont les tweets pertinents sont aussi récents. Nous allons discuter ce point en détail dans la section 5.6.3.3.

En se basant sur les scores interpolés des documents obtenus avec $\mu^{(1)}$ sur l'ensemble des requêtes de TREC 2011, nous avons calculé avec la méthode des moindres carrés (Cf., algorithme 1), la capacité à utiliser pour la phase de test. Cette capacité est composée de : (μ_{To} = 0.633, μ_{Re} = 0.204, μ_{Au} = 0.153, $\mu_{\{To,Re\}}$ = 0.961, $\mu_{\{To,Au\}}$ = −0.210, $\mu_{\{Re,Au\}}$ = −0.5). En effet, les critères topicalité et fraîcheur d'information se sont attribués les valeurs de capacités les plus élevées. Cette observation coïncide avec la nature de la tâche TREC Microblog, où les utilisateurs sont généralement intéressés aux tweets pertinents et récents. Nous remarquons aussi que les valeurs de capacités obtenues sur les sous ensembles $\{To, Au\}$ et $\{Au, Fr\}$ sont négatives. Selon les trois types d'interaction que nous avons présentés dans la section 5.4.1, la contribution du critère de pertinence topicalité à toute combinaison de critère contenant le critère autorité est supérieure à sa contribution quand le critère autorité est exclu. Il en est de même pour les deux critères fraîcheur des tweets et autorité. Ceci est probablement dû au fait que les assesseurs de la tâche Microblog de TREC n'ont pas donné une très grande importance à l'autorité des auteurs des tweets pendant leur évaluation de la collection de test. Pour interpréter ces résultats, nous présentons dans ce qui suit, les valeurs d'importance et d'interaction entre les différents critères.

5.6.3.2.1 Analyse de l'importance des critères de pertinence

Critère	*Topicalité*	*Fraîcheur d'information*	*Autorité*
Indice d'importance	**0.63**	**0.25**	**0.12**

Table 5.4: Indices d'importance des critères.

L'analyse d'importance des critères, présentée dans la figure 5.4, obtenue en utilisant l'indice d'importance montre une importance du critère thématique avec une valeur de 0.631. Le critère fraîcheur d'information est aussi donné une plus grande importance (0.25) comparé à l'autorité des utilisateur (0.12). Ceci peut s'expliquer par le fait que les requêtes de la tâche sur lesquelles nous avons effectué notre apprentissage des capacités, sont liées dans la plupart des cas à des documents thématiquement pertinents plutôt que récents ou autoritaires. Ainsi, lors de l'évaluation des tweets de la collection

par les assesseurs de TREC, ces derniers accordent plus d'importance à la correspondance thématique entre les requêtes et les documents.

5.6.3.2.2 Analyse de corrélation des critères de pertinence La Figure 5.5 montre les valeurs des indices d'interactions entre les trois critères thématique *To*, fraîcheur d'informations et autorité suivant les requêtes la tâche Microblog de TREC 2011. On remarque que le critère autorité n'est pas important et ne donne aucune contribution quand il est combiné avec les deux autres critères. Partant de ce constat, le critère autorité, malgré son importance relative dans Twitter (Chen *et al.*, 2012), peut biaiser les scores globaux des tweets, et ceci explique bien les capacités négatives assignées à $\mu_{\{To,Au\}}$ et $\mu_{\{Au,Fr\}}$. Cependant, nous notons une interaction positive entre le critère thématique et fraîcheur d'information, ce qui implique une contribution considérable aux scores globaux lorsqu'ils sont combinés ensemble. Ces valeurs conviennent à l'objectif de la tâche de recherche de tweets et au cadre de RI que nous avons considéré dans ce chapitre. De plus, étant donné que les utilisateurs de Twitter préfèrent les tweets qui sont à la fois pertinents et récents (*i.e.,* dont les scores sont balancés entre ces deux critères et non pas biaisés par l'un ou l'autre), l'intégrale de Choquet a répondu à cette préférence en associant une capacité élevée à la combinaison du sous ensemble de critères {topicalité, fraîcheur des tweets} par rapport aux scores individuels de chacun d'entre eux.

Critère	*Topicalité*	*Fraîcheur d'information*	*Autorité*
Topicalité	–	**+0.18**	**+0.01**
Fraîcheur	–	–	**-0.10**
Autorité	–	–	–

Table 5.5: Indices d'interaction des critères.

Pour vérifier la validité de ces résultats, nous présentons en plus des indices de Choquet, une analyse de corrélation des ordonnancements selon chaque critère de pertinence. Cette analyse, présentée dans la figure 5.6, est basée sur le coefficient de corrélation de Kendall (Kendall, 1938). Le *tau* de Kendall mesure la corrélation de rang entre deux listes d'ordonnancements. Elle est

donnée par :

$$tau = \frac{(nombredepairesconcordantes) - (nombredepairesdiscordantes)}{(1/2) \cdot n \cdot (n-1)} \tag{5.11}$$

Dans notre contexte, nous analysons la concordance entre les listes de documents retournés suivant chaque paire de critères de pertinence. Plus les listes sont similaires (resp., discordants), plus le coefficient *tau* est proche de 1 (resp., −1). Si les listes d'ordonnancements (i.e., les critères) sont indépendants, alors le coefficient est approximativement égal à zéro.

Critère	*Topicalité*	*Fraîcheur d'information*	*Autorité*
Topicalité	1	**0.1580**	**0.0013**
Fraîcheur	–	1	**0.0010**
Autorité	–	–	1

(a) Coefficient de corrélation de rang Kendall pour les critères individuels.

Critère	*{To, Fr}*	*{To, Au}*	*{Fr, Au}*
{To, Fr}	1	**0.2290**	**0.1210**
{To, Au}	–	1	**-0.1030**
{Fr, Au}	–	–	1

(b) Coefficient de corrélation de rang Kendall pour les sous ensembles critères.

TABLE 5.6: Analyse de corrélation des critères dans la collection des tweets.

Les tableaux (5.6a) et (5.6b) montrent les coefficent de corrélation des critères individuels ainsi que d'entre les sous ensembles de critères. Les résultats sont moyennés sur tous les documents de chaque critère. Le tableau 5.6a montre une corrélation significative entre les critères fraîcheur d'information et topicalité, tandis que le critère autorité semble être indépendant et moins important. Cependant, le tableau (5.6b) montre que ce dernier impacte les listes d'ordonnancements en présence des deux autres critères. Cet impact est plus significatif en présence du critère topicalité, ce qui n'est pas surprenant, vu les résultats déjà obtenu avec les indices d'interaction. Ces analyses sont en concordance avec les études menées par (Carterette *et al.*, 2011) et (Wei *et al.*, 2010) dans cette même direction de recherche.

5.6.3.3 Résultats expérimentaux

Dans cette section, nous analysons les résultats des performances de recherche de notre approche selon le protocole d'évaluation présenté dans la section 5.6.3.1.

5.6.3.3.1 Analyse avec les méthodes d'agrégation classiques et les opérateurs d'agrégation prioritaires Dans cette section, nous comparons notre modèle avec des méthodes de combinaison de scores classiques ainsi que les opérateurs prioritaires présentés dans la section 5.6.3.1. La méthode de combinaison linéaire est utilisée comme suit : $MCL(T) = \sum_{i=1}^{3}(\alpha_i mcl_i(T))$ où $mcl_i(T)$ est le score partiel d'un tweet selon un critère avec $i \in$ *{topicalité, autorité, fraîcheur}*. Les poids utilisés avec la méthode de combinaison linéaire sont empiriquement paramétrés dans la phase d'apprentissage, au sein des requêtes de la tâche Microblog 2011. Nous leur avons attribué les poids ayant donné les meilleurs valeurs en termes de précision $P@30$: $\alpha_{fraîcheur} = 0.23$, $\alpha_{autorité} = 0.16$ et $\alpha_{\text{topicalité}} = 0.61$, où α_i est le poids du critère c_i. Nous avons aussi paramétré l'opérateur d'agrégation prioritaire SCORING pour identifier le meilleur scénario de priorisation des critères, donnant ainsi résultat au scénario Sc_1 : {topicalité} ≻ {fraîcheur} ≻ {autorité}.

Le tableau 5.7 présente les résultats, en termes de $P@10$, $P@20$, $P@30$ et MAP obtenus par IAGGREGATOR ainsi que les référentiels de comparaison, sur les requêtes de la tâche Microblog de TREC 2012.

Comme le montre le tableau 5.7, IAGGREGATOR donne des résultats meilleurs que les autres méthodes d'agrégation, en termes des précision $P@30$ et de MAP. La significativité des éventuels accroissements obtenus ont été estimés en appliquant le test statistique de *student*. Les symboles †, ‡ et ⋆ marquent les différences significatives des valeurs de p obtenues. Nous remarquons à partir du tableau 5.7, que la différence de performance est plus importante en comparaison avec les méthodes d'agrégation classiques. Cette différence est de l'ordre de 60.26% avec la méthode WAM et 63.23% avec l'opérateur MAX. Pour l'opérateur prioritaire SCORING, la différence de performance est moins importante. Ainsi, le scénario de priorisation Sc_1 adopté favorise les tweets dont le critère thématique est plus satisfait que l'autorité et la fraîcheur, ce qui convient aussi aux préférences que nous avons trouvées sur les mesures capacités. La différence de performance peut être expliquée par la prise en compte d'autres poids d'importance entre les sous ensembles de critères, ce qui a permis d'atténuer les scores des tweets qui sa-

Opérateur	Précision				% ↗
	P@10	P@20	P@30	MAP	
Am	0.1140 ‡	0.0991 ⋆	0.0936 ⋆	0.0535	**+59.89%**
Wam	0.1161 ⋆	0.0991 ⋆	0.0929 ⋆	0.0539	**+60.28%**
Mcl	0.1860 ‡	0.1833 ‡	0.1854 ‡	0.0928	**+20,73%**
Max	0.1088 ⋆	0.0895 ⋆	0.0860 ⋆	0.0604	**+63,23%**
Min	0.1793 ‡	0.1767 ‡	0.1764 ⋆	0.0879	**+24.58%**
Owa	0.1879 †	0.1776 ‡	0.1764 ⋆	0.0882	**+24.58%**
OWMin	0.1897 †	0.1776 ‡	0.1833 ⋆	0.0902	**+21.63%**
And	0.1793 ‡	0.1767 ‡	0.1764 ⋆	0.0882	**+24.58%**
Scoring	0.2018 ‡	0.1982 ‡	0.1977 ⋆	0.1091	**+15.47%**
iAggregator	**0.2345**	**0.2293**	**0.2339**	**0.1252**	–
	+13.94%	**+13.56%**	**+15.47%**	**+12.85%**	

TABLE 5.7: Évaluation comparative des performances recherche. "% Amélioration" indique l'amélioration de notre approche en terme de $P@30$. Les symboles § et ⋆ dénotent le test t-*student* : "§" : $0.05 < t \leqslant 0.1$; "⋆" : $t \leqslant 0.01$.

tisfont uniquement les critères autorité et fraîcheur. Comparée à l'opérateur AND, les performances sont significativement meilleures. De plus, malgré la priorisation sur les critères, AND donne des résultats qui sont plus faibles que ceux obtenus par la méthode de combinaison linéaire MCL. Nous notons aussi que cette dernière donne des valeurs de précision qui sont aussi meilleurs que l'opérateur OWA. Ces observations peuvent être expliquées par le fait que OWA s'intéresse plutôt aux documents dont les scores sont élevés et donne ainsi des scores relativement faibles aux documents ne satisfaisant pas certains critères. Par conséquence, un poids faible sur un critère donné, peut être une raison majeure pour élaguer un document, ce qui peut donner des résultats biaisés par les scores les plus élevés. En ce qui concerne l'opérateur OWMIN, l'amélioration est de 20% qui est la même obtenue pour MCL. Cet opérateur utilise un vecteur pour représenter les niveaux d'importance, afin de minimiser l'impact des poids faibles des critères dans l'ordonnancement final des documents. A la différence de OWA, l'opérateur d'agrégation OWMIN se base sur l'opérateur *min* pour calculer les scores globaux des documents. A partir de cette analyse, nous pouvons conclure que parmi les raisons majeures des performances relativement faibles des ces méthodes est

le biais introduit par les documents dont les scores sont majoritairement satisfaits par un critère ou l'autre, d'autant plus que ces derniers présentent quelques dépendances comme nous l'avons montré dans la section 5.6.3.2.

Dans le but de donner un aperçu plus large des résultats obtenus dans le tableau 5.7, nous présentons dans la figure 5.7, les résultats des mesures de précision sur différents niveaux entre IAGGREGATOR et les référentiels de comparaison.

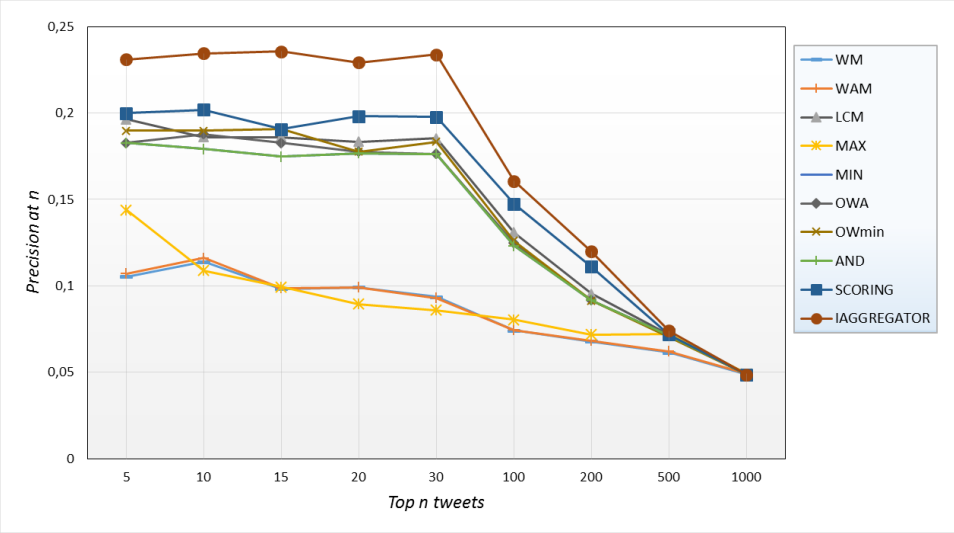


FIGURE 5.7: Précision à différents niveaux de coupe *@n* obtenues par IAGGREGATOR et les référentiels de comparaison.

Comme pour les résultats déjà présentés dans le tableau 5.7, les différences de performances en termes de précision entre notre modèle et les modèles de l'état de l'art sont plus importantes pour les opérateurs AM, WAM et l'opérateur MAX, sur les différents niveaux de coupe. Pour la méthode MAX, l'explication directe est que les scores globaux sont toujours dominés par les critères donnant les scores les plus élevés. Par exemple, le manque de satisfaction d'un critère comme la thématique de recherche, peut être compensé par la satisfaction du critère autorité, ce qui est non réaliste dans un contexte de recherche de tweets. Comme nous l'avons déjà discuté, pour les opérateurs MIN et AND, les résultats similaires obtenus peuvent être expliqués par le fait que le premier est généralement dominé par le score le plus faible alors que le deuxième pénalise les tweets qui satisfont le plus le critère le moins important. En effet, s'il existe beaucoup de tweets dont le critère

autorité est majoritairement satisfait par rapport aux deux autres critères, sa satisfaction globale pourrait être biaisée par ce critère de pertinence.

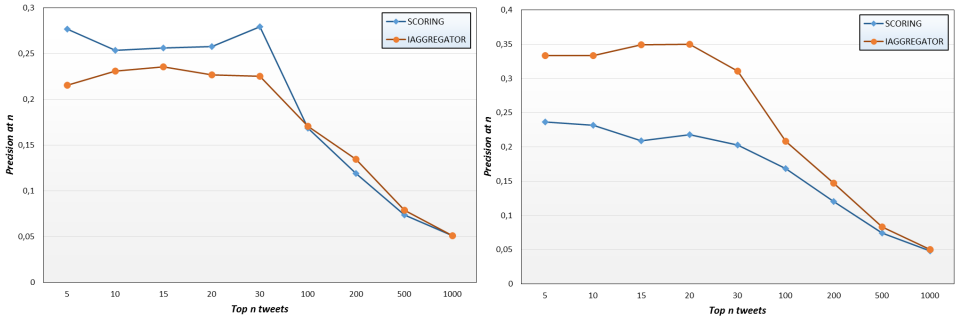
Nous présentons dans la suite, une analyse de défaillance de notre approche en comparaison avec les mêmes méthodes déjà présentées. Le tableau 5.8 présente le pourcentage des requêtes $\mathcal{R}^+$, $\mathcal{R}^-$ et $\mathcal{R}$ pour lesquels iAggregator est plus efficace (resp., moins performant, égale) les référentiels, en termes de $P@30$ avec une amélioration supérieure (resp., inférieure, égale) à 5% en comparaison avec les 5 meilleurs opérateurs.

Ensemble de requêtes	**Am**	**Min**	**Owa**	**Owmin**	**Scoring**
$\mathcal{R}^+$	$56,89\%$	$43,10\%$	$43,10\%$	$36,20\%$	$36,20\%$
$\mathcal{R}$	$22,41\%$	$37,93\%$	$37,93\%$	$43,10\%$	$41,37\%$
$\mathcal{R}^-$	$20,68\%$	$18,96\%$	$18,96\%$	$18,96\%$	$22,41\%$

Table 5.8: Pourcentage des requêtes $\mathcal{R}^+$, $\mathcal{R}^-$ et $\mathcal{R}$ pour lesquelles iAggregator est plus performant (resp., moins performant, égal) les référentiels, en termes de $P@30$.

Le tableau 5.8 montre que le pourcentage des requêtes pour lesquelles iAggregator est moins performant que les autres méthodes d'agrégation est presque le même, avec une moyenne de $20,34\%$. Une analyse manuelle des ces requêtes montre qu'elles sont pratiquement les mêmes pour tous les opérateurs, avec une petite différence pour la moyenne arithmétique Am. Le plus grand pourcentage pour les requêtes $\mathcal{R}^+$, est atteint également avec Am. La différence de pourcentage est aussi similaire pour les trois ensembles de requêtes suivant les méthodes d'agrégation de l'état de l'art. Nous notons aussi que le pourcentage le moins élevé pour l'ensemble $\mathcal{R}^+$ est marqué avec les deux opérateurs Scoring et OWmin, avec $36,2\%$ de requêtes, tandis que pour l'ensemble $\mathcal{R}^-$, la différence est également moins significative pour Scoring avec un taux de $22,41\%$.

Dans la figure (5.8), nous présentons les différences de performances en termes de $P@5 \ldots P@1000$, entre iAggregator et le meilleur référentiel Scoring pour les deux ensembles de requêtes $\mathcal{R}^+$ et $\mathcal{R}^-$. Comme nous pouvons le voir de la figure (5.8a), la différence de performance pour l'ensemble $\mathcal{R}^-$ n'est pas significative. Il est à noter aussi que notre opérateur donne des valeurs de $P@30$ nulles pour 4 requêtes de $\mathcal{R}^-$. La moyenne de différence de performance est environ $5,43\%$ et la meilleure amélioration est marquée



Comparaison des requêtes de l'ensemble $\mathcal{R}^-$ Comparaison des requêtes de l'ensemble $\mathcal{R}^+$

FIGURE 5.8: Précision à différents niveaux de coupe *@n* obtenues par IAGGREGATOR en comparaison avec l'opérateur d'agrégation prioritaire SCORING pour les deux ensembles $\mathcal{R}^-$ et $\mathcal{R}^+$ de requêtes.

pour n = 5 avec une différence de $22,21\%$. La différence la plus faible est observée pour les requêtes $T63$ et $T65$ des requêtes de la tâche TREC Microblog 2012 avec des valeurs de $75,01\%$ et $28,54\%$ respectivement. La première requête est : "*Bieber and Stewart trading places*" est une requête sensible au temps. Notre modèle n'a pas réussi à trouver les documents pertinents en premier lieu. Ceci peut être expliqué par le degré d'importance faible assigné au critère fraîcheur d'information (μ_{Re} = 0.204) en comparaison au critère topicalité (μ_{To} = 0.633), en dépit de la valeur de capacité un peu élevée attribuée à la combinaison des deux critères. Il en est de même pour la requête $T65$: "*Michelle Obama's obesity campaign*", et ceci est aussi dû à l'hypothèse que les tweets les plus pertinents sont ceux publiés dans des périodes de temps très proches du temps de soumission des requêtes. Donc, si on suppose que les tweets les plus récents doivent avoir les scores les plus élevés, ceci peut affecter la position des tweets pertinents dans la liste d'ordonnancements s'ils sont pas récemment publiés. Une idée intéressante dans ce cas de figure, pourrait consister à reparamétrer les valeurs des capacités en fonction du type de la requête et sa sensibilité au temps.

Pour les requêtes de l'ensemble $\mathcal{R}^+$, pour lesquelles IAGGREGATOR améliore SCORING, nous pouvons voir à partir de la figure (5.8b), que la différence de performance est significative. Cette différence est plus significative pour les premiers *top* 30 tweets avec une moyenne de $34,41\%$, contrairement à l'ensemble $\mathcal{R}^-$ $14,10\%$ pour le même ensemble de tweets. Considérons par exemple la requête 73 : "*Iran nuclear program*", nous remarquons que IAGGREGATOR donne des bons résultats par rapport au référentiel de compa-

raison. Cette requête présente plusieurs fenêtres temporelles pertinentes, à la différence des deux requêtes $T63$ et $T65$, pour lesquelles les tweets pertinents sont dans un intervalle de temps très précis (i.e., proche du temps de soumission de la requête).

5.6.3.3.2 Évaluation avec les algorithmes d'apprentissage d'ordonnancements Dans cette section nous présentons une comparaison de notre modèle avec les méthodes d'apprentissage d'ordonnancements présentées dans la section 5.6.3.1. Plus spécifiquement, nous testons iAggregator aux méthodes RankNet, RankSVM, ListNet, Random Forest (RF) et λ–MART. Nous avons utilisé l'outil open source pour paramétrer RankSVM Joachims (2006) et nous avons exploité la librairie RankLib pour les algorithmes RankNet et ListNet [11]. Pour tous les réglages, nous avons utilisé 200 itérations avec comme la mesure P@30 comme fonction de perte. Les modèles sont appris sur le même ensemble d'apprentissage utilisé pour le paramétrage des valeurs de capacités (Cf., 5.6.3.2). Ces expérimentations sont effectuées avec une validation croisée avec un $k = 5$.

Opérateur	**Précision**				% ↗
	P@10	**P@20**	**P@30**	**MAP**	
RankSVM	0.2500 ‡	0.2250 †	0.2218 †	0.0871	**+5.17%**
RankNet	0.2448 †	0.2198 †	0.2201 †	0.0858	**+5.89%**
ListNet	0.0931 ⋆	0.1009 ⋆	0.1115 ⋆	0.0485	**+52.33%**
RF	0.0810	0.0681	0.0687	0.0628	**+70,68%** ⋆
λ–MART	0.2276	0.2092	0.2043	0.1856	**+11,67%** ⋆
iAggregator	**0.2345**	**0.2293**	**0.2339**	**0.1252**	–
	-6.60%	**+1.87%**	**+5.17%**	**+30.43%**	

Table 5.9: Évaluation comparative des performances recherche de notre méthode avec les algorithmes d'apprentissage d'ordonnancement. La dernière ligne indique la différence de précision avec RankSvm.

Le Tableau 5.9 montre que notre approche donne de bons résultats en comparaison avec les algorithmes d'apprentissage d'ordonnancements. Le taux d'amélioration est plus important pour l'algorithme RF avec un pourcentage de $70,68\%$, alors qu'il est beaucoup moins élevé pour les algorithmes

11. `http://people.cs.umass.edu/~vdang/ranklib.html`

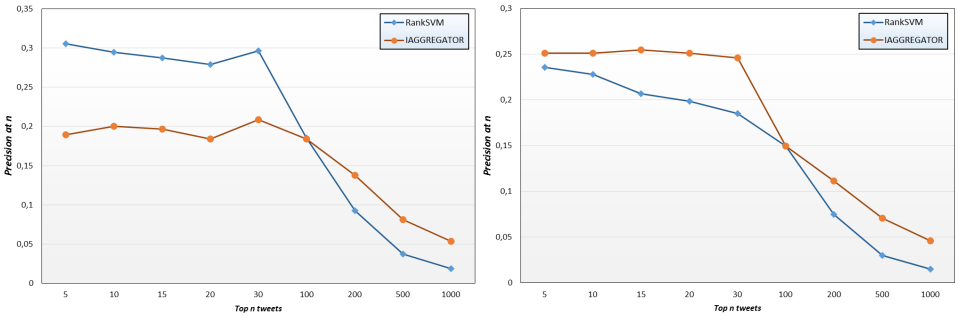
RankSVM et RankNet. En effet, les performances de RankSVM n'est pas surprenante sur la recherche de tweets, vu les résultats donnés dans des études antérieures Duan *et al.* (2010). Cependant, ces améliorations sont plus élevées en termes de la mesure MAP, avec une valeur de 30.43%.

Pour donner une analyse plus approfondie des améliorations de iAggregator en comparaison avec les algorithmes d'apprentissage d'ordonnancement, nous présentons dans la suite une analyse de défaillance avec RankSVM et RankNet. Le tableau 5.10 montre les pourcentages des requêtes $\mathcal{R}^+$ et $\mathcal{R}^-$ pour lesquelles iAggregator donne des meilleurs (resp., plus faibles) résultats en termes de $P@30$.

Ensemble des requêtes	**RankSVM**	**RankNet**
$\mathcal{R}^+$	$67,24\%$	$67,24\%$
$\mathcal{R}^-$	$32,76\%$	$32,76\%$

Table 5.10: Pourcentage des requêtes $\mathcal{R}^+$ et $\mathcal{R}^-$ pour lesquelles iAggregator donne des meilleurs (resp., plus faibles) résultats en termes de $P@30$.

Nous remarquons que le pourcentage des requêtes de $\mathcal{R}^+$ est environ $67,24\%$ pour les deux requêtes. Malgré les pourcentages similaires obtenus pour $\mathcal{R}^+$ et $\mathcal{R}^-$, l'analyse de ces requêtes a montré qu'elles sont pas totalement les mêmes pour les deux algorithmes.



Comparaison pour les requêtes de l'ensemble $\mathcal{R}^-$ — Comparaison pour les requêtes de l'ensemble $\mathcal{R}^+$

Figure 5.9: Précision à différents niveaux de coupe $@n$ obtenues par iAggregator en comparaison avec RankSVM pour les deux ensembles de requêtes $\mathcal{R}^-$ et $\mathcal{R}^+$.

Dans la figure (5.9), nous présentons les différences de performances de iAggregator et RankSVM en termes de $P@5 \ldots P@1000$ pour les deux ensembles $\mathcal{R}^+$, $\mathcal{R}^-$. Nous notons à partir de la figure que l'amélioration est plus significative pour les requêtes de l'ensemble $\mathcal{R}^-$. Ceci n'est pas surprenant étant donné que le pourcentage des requêtes pour lesquelles iAggregator améliore RankSVM est relativement élevé (jusqu'à $67,24\%$), malgré le fait que l'amélioration en termes de $P@30$ est nettement plus faible (+5.17%). Pour l'ensemble $\mathcal{R}^-$, nous remarquons que RankSVM améliore iAggregator uniquement pour les premiers 100 tweets.

5.6.3.3.3 Comparaison avec les résultats officiels de la tâche TREC Microblog Dans la suite, nous comparons nos résultats avec les résultats officiels des participants à la tâche Microblog de TREC 2012 en termes de $P@30$ et MAP.

Modèle	**P@30**	**MAP**
Meilleur *run*	0.2701	0.2642
Second Meilleur *run*	0.2559	0.2277
median run	0.1808	0.1480
iAggregator	0.2339	0.1252

Table 5.11: Comparaison avec les résultats officiels des participants à la tâche Microblog de TREC 2012 en termes de $P@30$ et MAP.

Les résultats du tableau 5.11 montrent que notre modèle est meilleur que le *median run* de la tâche en termes de $P@30$ et MAP, i.e., la valeur médiane des deux mesures calculées sur tous les systèmes participants à la tâche. Il est à noter que nous avons uniquement exploité trois critères de pertinence, parmi plusieurs facteurs, étant donné que notre objectif n'est pas la recherche de tweets en soi, mais plutôt l'optimisation de la fonction de la fonction de combinaison de scores quelque soit les critères utilisés. De plus, mis à part l'apprentissage que nous avons effectué pour paramétrer les valeurs des mesures floues ainsi que les paramètres des référentiels de comparaison, nous n'avons utilisé aucune source d'évidence externe, telle que Wikipedia ou des ontologies. Les différences de performances que nous avons déjà présentées montrent que les performances faibles ne sont pas liées au processus d'agrégation, mais plutôt au processus de sélection initial des

documents pertinents. Ceci pourrait aussi s'expliquer par le fait que nous n'avons pas participé à la tâche, ce qui fait que les tweets que nous retournons, n'ont pas été jugés par les assesseurs de NIST, et n'ont pas donc fait partie du pool. Tout de même, nous assumons que les résultats donnés par notre modèle sont satisfaisants.

5.6.4 Évaluation de iAggregator dans un cadre de RI contextuelle

Dans cette section, nous évaluons notre modèle d'agrégation personnalisée dans le cadre d'une tâche de recherche de lieux d'attraction. Plus spécifiquement, nous exploitons les requêtes et les données fournies par la tâche TREC *Contextual Suggestion* 2013 (Cf., section 5.6.2.2).

5.6.4.1 Protocole d'évaluation

Comme dans le cas de la tâche de recherche de tweets, nous avons adopté ici une méthodologie entièrement automatisée afin d'identifier les valeurs des capacités pour tous les utilisateurs et tester les performances du modèle d'agrégation. À cette fin, nous avons procédé à une partition aléatoire de l'ensemble des 50 contextes en deux sous ensembles de même taille, noté Q^{app} et Q^{test} utilisés respectivement pour l'apprentissage et le test. En outre, pour éviter le problème de sur-apprentissage, l'ensemble des contextes est divisé aléatoirement dans un second tour en deux ensembles différents d'apprentissage et de test.

Enfin, pour tester l'efficacité de notre modèle, nous nous sommes appuyés sur l'ensemble de contextes restants Q^{test}, et nous avons utilisé la mesure officielle $P@5$ pour le calcul de performances. Cette mesure de précision est équivalente à la proportion des suggestions de lieux pertinents retournés parmi les 5 premiers résultats.

5.6.4.2 Paramétrage et identification des mesures floues

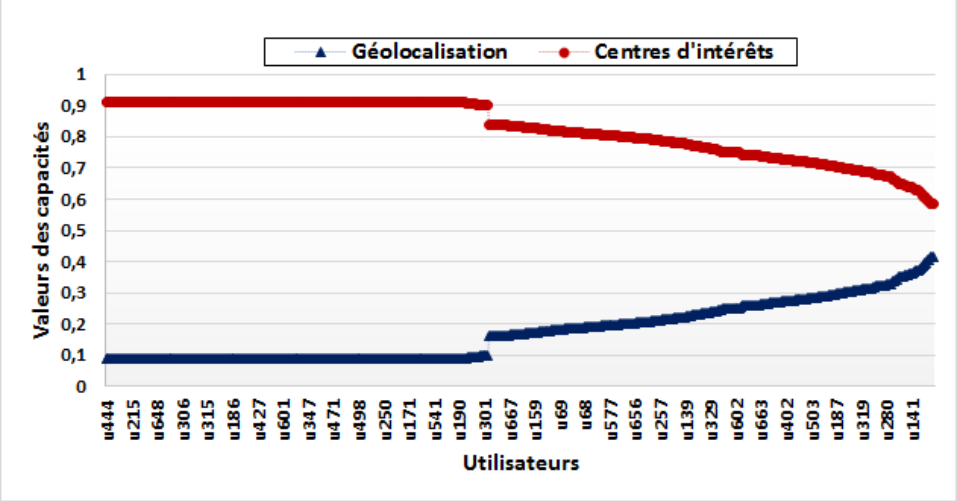
L'objectif principal de la phase d'apprentissage est d'apprendre les capacités ($\mu^u_{\{centre\_interet\}}$, $\mu^u_{\{localisation\}}$) qui correspondent à l'importance des deux critères de pertinence. Nous commençons d'abord par une mesure floue initiale donnant le même poids d'importance pour les deux critères de perti-

nence. Ensuite, nous calculons la mesure de précision $P@5$ (mesure officielle de la tâche) de tous les contextes de l'ensemble d'apprentissage Q^{app}. En utilisant la vérité de terrain fournie avec la tâche "Contextual Suggestion" de TREC 2013, et en se basant sur l'algorithme 1, nous identifions pour chaque utilisateur ses préférences sur les deux critères : centres d'intérêts et localisation géographique. Les différentes valeurs obtenues pour chaque utilisateur sont données dans la suite.

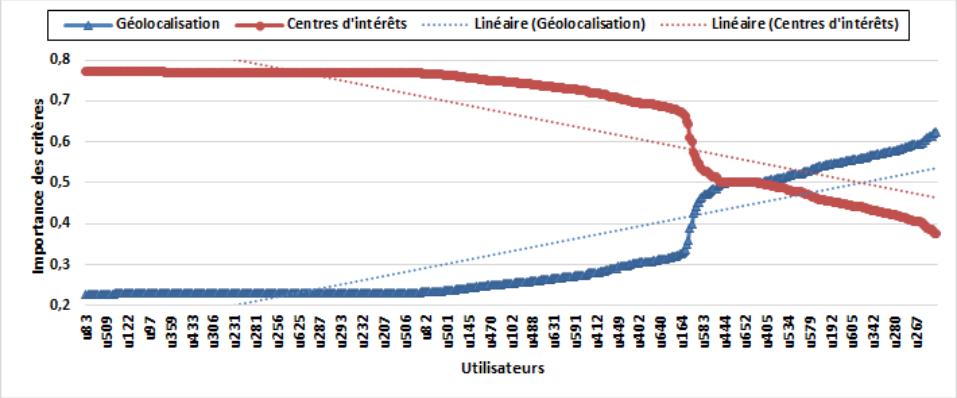
5.6.4.2.1 Analyse de l'importance des critères de pertinence Dans cette section, nous analysons l'importance des critères pour les utilisateurs ($\mu^u_{\{centre\_interet\}}$, $\mu^u_{\{geolocalisation\}}$) pour chaque utilisateur (u). A cet effet, nous commençons par analyser l'importance intrinsèque de chaque critère indépendamment de l'autre critère. La Figure 5.10a montre la variation des valeurs de capacité pour chaque utilisateur selon les deux critères de pertinence sur l'ensemble Q^{app} d'apprentissage. L'axe des abscisses représente l'ensemble des utilisateurs (35-669) et l'axe des ordonnées représente les valeurs des capacités correspondantes selon les critères centres d'intérêt (Ci) et géolocalisation (Geo).

En se référant à la Figure 5.10a, nous constatons que le critère Ci se voit accorder une capacité plus importante que le critère Geo. Par exemple, l'utilisateur 285 a une valeur de capacité de l'ordre de $0,23$ pour le premier critère alors qu'il a une mesure de l'ordre de $0,76$ pour le critère Geo. Ceci est prévisible étant donné que les utilisateurs de cette tâche s'intéressent généralement aux lieux qui correspondent principalement à leurs préférences, même si elles ne sont pas géographiquement pertinentes. Cependant, la Figure (5.10a) montre que la distribution des valeurs de capacité est loin d'être la même pour tous les utilisateurs et met en exergue des valeurs qui vont de $0,09$ à $0,414$ pour le critère Geo et d'autres qui vont de $0,585$ à $0,909$ pour le critère Ci.

Pour mieux comprendre ce constat, nous traçons sur la Figure 5.10b, les valeurs des indices d'importance reflétant, pour chaque utilisateur, le degré de préférence globale selon les deux critères de pertinence Ci et Geo. A la différence de la Figure 5.10a, la Figure 5.10b met en évidence l'importance moyenne de chaque critère de pertinence quand il est associé à l'autre critère. On peut observer sur la Figure 5.10b que les préférences des utilisateurs sur les deux critères sont totalement différentes. Le lissage



(a) Valeurs de capacités des utilisateurs suivant les deux critères de pertinence centres d'intérêt et géolocalisation.



(b) Importance des critères centres d'intérêt (Ci) et géolocalisation (Geo).

FIGURE 5.10: Valeurs de capacités des utilisateurs et importance des critères de la tâche "Contextual Suggestion" de TREC 2013.

des valeurs d'importance obtenues selon ces critères donne deux courbes linéaires avec des valeurs tout à fait constantes et différentes, corroborant ainsi les résultats obtenus sur la Figure 5.10a. Le critère "centre d'intérêt" est encore pondéré par une importance relativement élevée pour la plupart des utilisateurs. Néanmoins, on peut également remarquer au milieu de la figure (valeurs comprises entre $0,4$ et $0,7$) que certains utilisateurs ont une préférence élevée sur le critère géolocalisation et inversement.

5.6.4.2.2 Analyse de corrélation des critères de pertinence Dans une seconde étape, nous analysons à travers la Figure 5.11, la dépendance entre les critères pour chaque utilisateur par le biais de l'indice d'interaction (Grabisch, 1995).

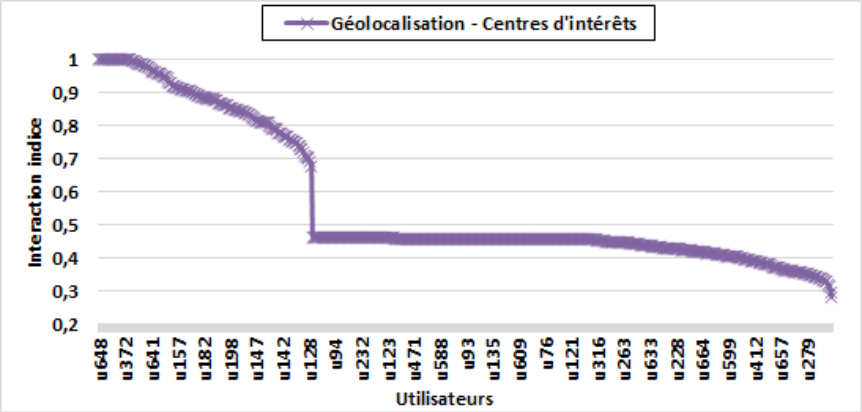


FIGURE 5.11: Indices d'intéraction entre les critères de pertinence centres d'intérêt et géolocalisation pour chaque utilisateur.

Plus les valeurs de cet indice sont proches de 1 (*resp.,* −1) plus les deux critères sont dépendants et l'interaction est positive (*resp.,* négative). Si la valeur de l'indice d'interaction est égale 0, les deux critères sont considérés comme indépendants et par conséquent, il n'existe aucune interaction entre ces derniers. On peut constater que les valeurs obtenues sur tous les utilisateurs sont toutes positives et varient entre $0,28$ et $0,99$. La valeur moyenne est de l'ordre de $0,56$ ce qui implique une interaction positive entre les deux critères de pertinence considérés lorsqu'ils sont combinés ensemble.

5.6.4.3 Résultats expérimentaux

Notre second objectif est d'évaluer l'efficacité de notre approche en termes :
- (*i*) d'agrégation de pertinence multidimensionnelle ;
- (*ii*) de personnalisation des préférences des utilisateurs sur les critères de pertinence.

Pour ce faire, nous comparons les résultats obtenus sur l'ensemble de contextes de test Q^{test} aux méthodes d'agrégation de référence (*baseline)* : la moyenne arithmétique pondérée (MAP) largement utilisée dans la plupart des approches impliquant la combinaison des scores de pertinence et les

deux opérateurs d'agrégation prioritaires Scoring et And, précédemment utilisés pour l'agrégation de pertinence dans un cadre de RI personnalisée (da Costa Pereira *et al.*, 2012). Il convient de préciser que nous avons effectué une série d'expérimentations avec une validation croisée pour identifier les meilleurs scénarios de priorisation devant être utilisés avec les deux opérateurs Scoring et And sur le même ensemble d'apprentissage utilisé pour trouver les valeurs de capacité de Choquet. Comme pour les résultats obtenus dans la phase d'analyse des indices d'importance, nous avons également constaté que le meilleur scénario est celui donnant une priorité au critère "centres d'intérêt" des utilisateurs. Cependant, les opérateurs d'agrégation ne sont pas en mesure de quantifier le degré d'importance des critères comme c'est le cas pour l'intégrale de Choquet.
Afin de montrer l'efficacité de l'approche de personnalisation, nous comparons l'opérateur d'agrégation personnalisé *versus* l'opérateur Choquet classique non personnalisé. Les capacités utilisées avec l'opérateur de Choquet classique sont obtenus en appliquant l'algorithme 1 une seule fois (et non pas pour chaque utilisateur), donnant ainsi en sortie des valeurs d'importance sur les critères indépendamment des préférences individuelles de chaque utilisateur. Ceci donne lieu à une valeur de $0,86$ pour le critère centre d'intérêt et une valeur de l'ordre de $0,14$ pour le critère géolocalisation. Les mesures de précision obtenues sont moyennées sur toutes les séries de tests et pour tout l'ensemble des requêtes de test.

La Figure 5.12 présente les résultats obtenus par notre approche, en comparaison avec les méthodes de référence. La Figure 5.12 montre que les performances de notre modèle sont plus élevées que les autres méthodes suivant la mesure officielle $P@5$, et également suivant les autres mesures.

La meilleure amélioration obtenue par notre approche suivant $P@5$ est marquée avec la méthode WAM (13.98%). En comparaison avec la meilleure méthode de référence (*i.e.,* And), les améliorations sont significatives mais moins importantes (10.11%) en termes de $P@5$. Ces résultats sont probablement dus au fait que l'opérateur d'agrégation prioritaire And est principalement basé sur l'opérateur Min, ceci pourrait pénaliser les lieux pertinents selon le critère le moins important à savoir, le critère géolocalisation. Vu que la plupart des utilisateurs ont une préférence moins importante selon ce critère, la pénalisation de ce dernier permet d'améliorer les performances de recherche. La différence obtenue dans la performance, en faveur de notre modèle, s'explique par la prise en compte des différents niveaux de préférence suivant les deux critères de pertinence ainsi que la prise en compte de l'interaction qui existe entre ces derniers.

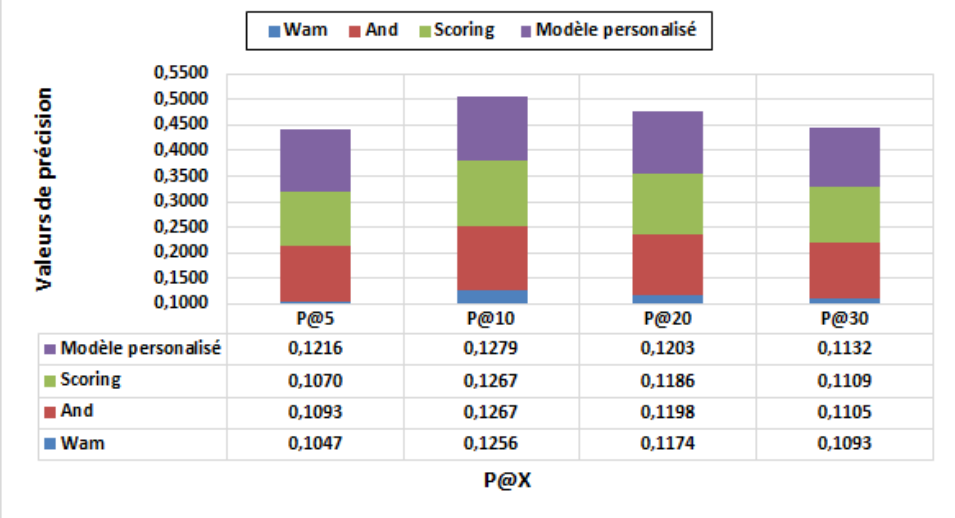


FIGURE 5.12: Efficacité de notre approche d'agrégation de pertinence dans la tâche "Contextual Suggestion" de TREC 2013 en comparaison avec les méthodes de référence.

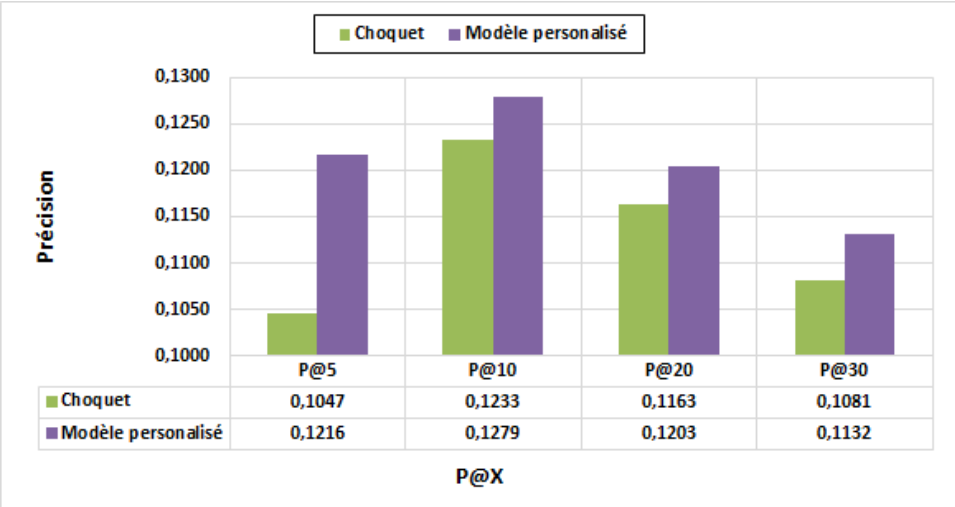


FIGURE 5.13: Efficacité de notre approche en terme de personnalisation en comparaison avec l'opérateur d'agrégation de Choquet classique.

En termes de personnalisation, la Figure 5.13 présente les résultats obtenus en termes de précisions (P@5, P@10, P@20 et P@30) entre l'opérateur

classique Choquet et sa version personnalisée. Ces résultats montrent que le dernier donne des plus bons résultats sur les différents niveaux précision. La meilleure amélioration est de l'ordre de $9,29\%$ en termes de $P@5$. Ces résultats confirment ceux obtenus dans la phase d'identification des capacités (Cf. section 5.6.4.2.1) où nous avons montré que les degrés d'importance des critères dépendent des préférences de l'utilisateur et ne sont pas les mêmes pour tous. La prise en compte des poids d'importance appropriés pour chaque critère et chaque utilisateur permet de donner ainsi des résultats à la fois pertinents et relativement adaptés aux préférences personnelles des utilisateurs.

5.6.4.3.1 Comparaison avec les résultats officiels de la tâche TREC Contextual Suggestion Dans cette section, nous présentons nos résultats obtenus avec l'opérateur de Choquet dans le cadre de notre participation à la tâche *Contextual Suggestion* de TREC 2014. Nous présentons également les meilleurs résultats obtenus par les autres groupes participants, en termes de $P@5$, TBG et MRR. En 2014, 17 groupes ont participé avec 31 soumissions (*run*). Pour notre part, nous avons participé avec une seule soumission en utilisant notre modèle se basant sur l'intégrale de Choquet. Le tableau 5.12 montre les valeurs de la mesure officielle $P@5$ utilisée obtenues par le meilleur système ainsi que notre modèle. Le *median run* donne une idée sur les valeurs médianes des résultats de tous les 31 systèmes ayant participé à la tâche.

Modèle	**P@5**	**TBG**	**MRR**
Meilleur *run*	0.5585	2.7021	0.7482
Second Meilleur *run*	0.5017	2.3718	0.6846
median run	0.4167	1.7684	0.5916
Opérateur de Choquet	0.2194	0.3331	0.3412

Table 5.12: Comparaison avec les résultats officiels de notre système participant à la tâche *Contextual Suggestion* de TREC 2014 avec les autres groupes participants, en termes de $P@5$, TBG et MRR. Le *median run* représente les valeurs médianes des résultats de tous les 31 systèmes ayant participé à la tâche.

On voit clairement à partir du tableau 5.12 que notre système est loin de la médiane ainsi que du meilleur système. Sur la totalité des systèmes soumis, 12 autres sont au dessous de la médiane. Notre seule soumission a été classée 19 sur les 31 autres soumissions à la tâche.

Néanmoins, nous notons que avons commencé par une approche très basique pour la recherche de lieux d'attraction. Ensuite, nous avons calculé les scores des deux critères exploités et nous les avons combiné avec notre modèle. Il s'avère enfin que cette approche est insuffisante, vu la taille faible de l'ensemble d'apprentissage, pour ce problème de personnalisation.

5.6.5 Évaluation de iAggregator dans un cadre de RI dans les folksonomies

5.6.5.1 Protocole d'évaluation

Pour cette troisième tâche de recherche, nous utilisons 75% des requêtes pour l'apprentissage et nous exploitons le reste des requêtes pour le test. Étant donné que pour cette collection de documents, nous disposons uniquement des 5 premiers résultats pertinents pour chaque requête, nous exploitons la mesure $P@5$ pour l'évaluation des résultats de recherche comme recommandé dans (Vallet et Castells, 2012).

5.6.5.2 Paramétrage et identification des mesures floues

Nous avons effectué le paramétrage des mesures floues des deux critères de cette tâche comme pour la tâche de recherche de tweets et la tâche de recherche de RI contextuelle. La figure 5.14 montre que les valeurs de précision sont presque les mêmes pour toutes les valeurs que nous avons testées. Les trois combinaisons de capacités marquées sur la figure (avec des cercles noires), montrent que les degrés d'importance n'influent pas sur les deux critères. Favoriser le critère centre d'intérêt (Us) ou le critère thématique (To) n'a pas d'impact sur la précision $P@5$. Nous allons analyser cette observation dans la section suivante.

5.6.5.2.1 Analyse de l'importance et de corrélation des critères de pertinence Après le calcul des indices d'importance des deux critères utilisés dans cette tâche, nous avons trouvé que tous les deux ont presque

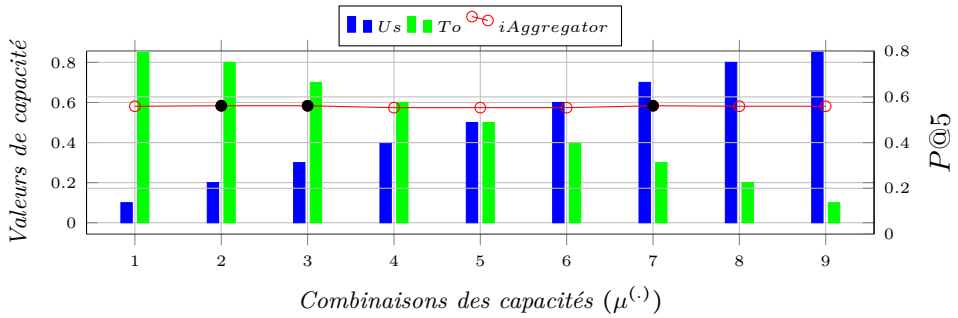


FIGURE 5.14: Résultats de précisions pour les valeurs de capacités à paramétrer dans la tâche de RI contextuelle. Ces valeurs sont obtenus sans personnalisation.

le même degré d'importance avec une valeur de 0.48 pour la pertinence thématique et 0.51 pour la pertinence utilisateur. Par ailleurs, la valeur d'interaction entre ces deux derniers est égale à 0.028, ce qui est un peu faible pour supposer qu'ils sont réellement dépendants. Ce résultat a eu un effet sur les performances de notre approche (cf. section 5.6.5.3).

5.6.5.3 Résultats expérimentaux

Dans cette section, nous évaluons notre approche d'agrégation personnalisée dans le cadre de la RI dans les folksonomies. Dans le Tableau 5.13, nous remarquons que les résultats obtenus avec l'opérateur de Choquet sont très proches que les résultats obtenus avec les modèles de référence. Ceci peut être expliqué par le fait que le nombre de critères est réduit d'une part et qu'ils sont en plus indépendants d'autre part, comme montré dans la section 5.6.5.2.1.

	MCL	**OWA**	**AND**	**SCORING**	**RankSVM**	**Modèle**
P@5	**0.6310**	**0.6310**	0.6286	**0.6310**	0.6286	**0.6310**
% ↗	0%	0%	**+0.003%**	0%	**+0.003%**	_
	⋆ ⋆ ⋆	⋆ ⋆ ⋆	⋆ ⋆ ⋆	⋆ ⋆ ⋆	⋆ ⋆ ⋆	

TABLE 5.13: Evaluation comparative des performances de recherche dans le contexte de RI personnalisée. Le symbole "⋆" dénote le test t−student : "⋆ ⋆ ⋆" : $t \leqslant 0.01$.

5.7 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche pour l'agrégation de pertinence multidimensionnelle. En se basant sur l'intégrale de Choquet et sur le concept de mesure floue, l'approche proposée est capable de modéliser les interactions pouvant exister entre les critères de pertinence. En effet, nous avons effectué un apprentissage des mesures floues, donnant lieu ainsi à une méthode générique qui est capable de traiter le problème d'agrégation multicritères indépendamment du nombre de critères utilisés et quelque soit le cadre d'agrégation. L'évaluation de notre approche, dans une tâche de recherche de tweets et sur les collection de test fournie par les tâches Microblog de TREC 2011, et 2012 a montré des résultats encourageants par rapport aux méthodes d'agrégation standards et quelques méthodes d'apprentissage d'ordonnancement.

Dans la deuxième partie de ce chapitre, nous avons tenté de personnaliser la méthode proposée afin de pondérer les préférences des utilisateurs à chacun des critères agrégés. Nous avons donc évalué l'impact de cette personnalisation dans deux tâches de recherche de RI personnalisée. Les résultats ont montré que le nombre de critères moins élevé a un effet sur les performances de recherche en comparaison avec des méthodes d'agrégation standards.

Chapitre 6

Vers une approche d'agrégation guidée par la requête : évaluation dans le cadre d'une tâche de RI sensible au temps

---

6.1 Introduction

Dans ce chapitre, nous nous intéressons à l'intégration de la dimension de pertinence temporelle dans le processus global d'agrégation multicritères. Ainsi, nous proposons une méthode d'agrégation multidimensionnelle permettant de combiner le critère temporel avec le critère thématique pour répondre à des besoins d'utilisateurs dépendant du temps. Nous modélisons cette tâche comme un problème d'agrégation d'ordonnancements. Ainsi, nous considérons que les termes de la requête sont temporellement corrélés et donc, nous supposons que chaque terme peut correspondre à une requête à part. Partant de ce postulat, nous fusionnons les listes des résultats issues de chaque terme pour constituer une seule liste agrégée qui correspond à l'ordonnancement final des documents. En effet, les principales contributions dans ce chapitre sont :

– Une analyse de la distribution temporelle des mots des requêtes qui montre

leur dépendance (temporelle), et ce au sein d'une large collection de documents qui évolue au cours du temps. Cette analyse est motivée par l'intuition que les documents pertinents sont ceux qui sont publiés dans les périodes de temps *rafales* et qui sont *bien* classés dans les listes des résultats de tous (ou la majorité) les termes d'une même requête.
– Une approche d'ordonnancement sensible au temps basée sur un modèle de langue temporel existant (Li et Croft, 2003; Efron et Golovchinsky, 2011; Dakka *et al.*, 2012) fournissant un cadre adéquat pour la combinaison des critères (Moulahi *et al.*, 2015b). Nous nous basons sur une méthode d'agrégation d'ordonnancements (Cormack *et al.*, 2009) pour formuler la tâche comme un problème de fusion de données, où chaque terme est considérée comme une requête en soi (Aslam et Montague, 2001). Cette formulation permet de *booster* les documents qui sont publiés dans les mêmes périodes de temps qu'un grand nombre d'autres documents pertinents.
– Une évaluation expérimentale basée sur une large collection de données standard dédiée aux tests des méthodes de RI sensibles au temps. Cette collection est fournie par les tâches *Knowledge Base Acceleration* (KBA) et *Temporal Summarization* (TS) de TREC 2013 et 2014.

La suite de ce chapitre est organisée comme suit. La section 6.2 présente nos motivations et quelques questions de recherche. Dans la section 6.3, nous introduisons notre modèle d'agrégation d'ordonnancements sensible au temps. Nous formalisons le problème et nous détaillons également toutes les étapes de l'approche proposée. L'évaluation expérimentale est présentée dans la section 6.4. Enfin, la section 6.5 conclut le chapitre.

6.2 Motivations et questions de recherche

Dans les deux dernières décennies, de nombreux études et revues ont été publiées dans le domaine de RI temporelle pour souligner l'importance des signaux temporels, surtout dans les collections traitant des documents liés aux actualités (Berberich *et al.*, 2010; Joho *et al.*, 2013; Campos *et al.*, 2014a; Wei *et al.*, 2014). Plusieurs modèles de RI ont exploité la dimension temporelle pour améliorer la qualité des SRI (Dong *et al.*, 2010; Efron et Golovchinsky, 2011; Massoudi *et al.*, 2011; Li et Croft, 2003). En dépit de la performance de ces modèles, ces derniers n'exploitent pas complètement les informations temporelles contenues dans les documents et les requêtes. La plupart des méthodes existantes se basent sur le modèle de langue basique ou une combinaison linéaire simple du facteur temporel avec le facteur

thématique où les termes de la requête sont supposés être générés indépendamment les uns des autres. Cependant, ces termes peuvent présenter une sorte de corrélation dans le temps, et cette corrélation peut être un indicateur temporel très important dans un modèle de RI. Examiner la distribution temporelle de chaque critère à part permet d'identifier les périodes de temps intéressantes pour chaque requête. Par exemple, les internautes ont tendance à parler de la "*fifa world cup*" principalement durant ou un peu après les périodes de temps où le tournoi a eu lieu. Par conséquent, les documents crées en dehors de ces périodes de temps sont moins susceptibles de discuter le tournoi en question, même s'ils contiennent un des termes de la requête, et donc ils ne sont peut être pas pertinents.

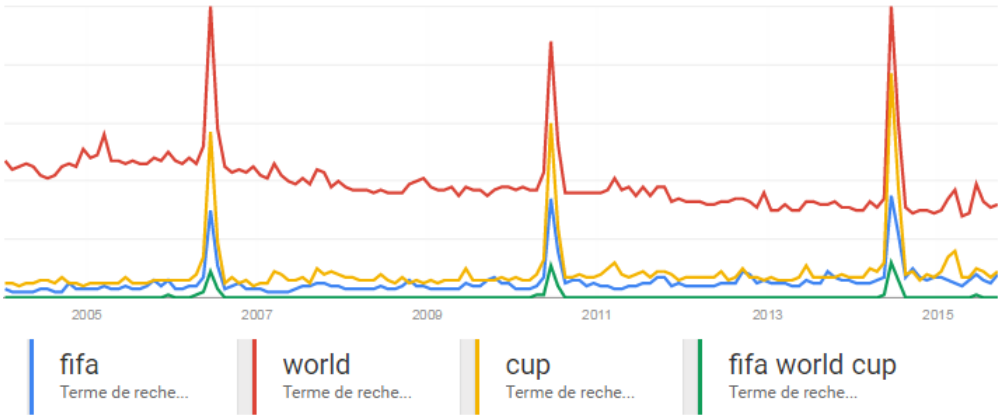


FIGURE 6.1: Évolution de l'intérêt des termes "fifa", "world", "cup" et "fifa world cup", au cours du temps à partir de *Google Trend* (accès en Septembre 2015). La distribution temporelle est donnée sur les 10 dernières années.

La figure 6.1 illustre bien cette observation. Elle montre l'évolution dans le temps de l'intérêt des termes "*fifa*", "*world*", "*cup*" et "*fifa world cup*" comme donnée par *Google Trend*[1]. La figure montre clairement, à travers les rafales, la corrélation temporelle entre les trois termes de la requête ainsi que la requête entière dans des périodes de temps bien particulières (e.g., Juillet 2010 et 2014). Ceci constitue notre intuition que les documents qui sont susceptibles d'être pertinents en réponse à une requête, sont ceux qui sont à la fois pertinents pour tous les termes de la requête et qui sont publiés dans des périodes de temps similaires.

1. `https://www.google.com/trends/`

Pour vérifier cette hypothèse dans des collections de données du monde réel, nous présentons dans la figure 6.2, une analyse des séries temporelles (Montgomery *et al.*, 2008) des documents pertinents de 6 requêtes ($Q1 - Q6$) de la collection *Stream Corpus* utilisée par la tâche Temporal Summarization [2] de TREC 2013. Dans la figure 6.2, l'axe des abscisses représente le temps en heures (l'age des documents obtenu par la différence entre le temps de soumission de la requête et l'estampille du document), et l'axe des ordonnées indique le poids normalisé de l'importance de la requête et des termes dans les documents pertinents.

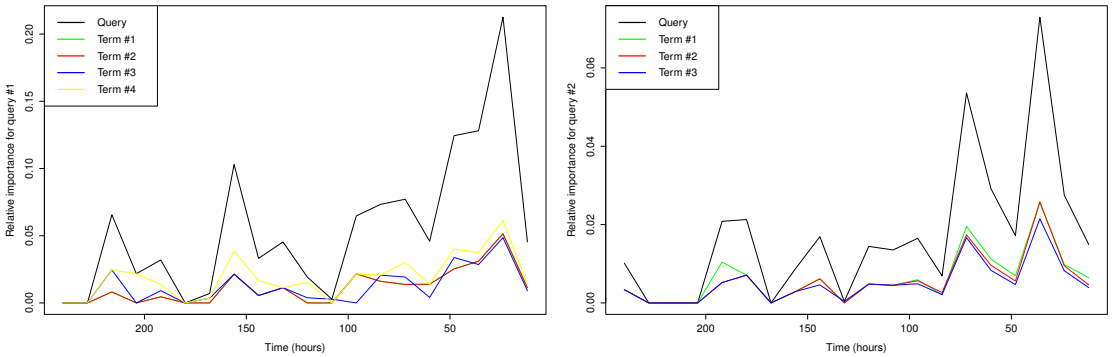
La figure montre que la distribution temporelle de la plupart des documents pertinents est relativement la même pour les termes et les requêtes les contenant. Cette observation renforce notre hypothèse de corrélation temporelle que nous avons déjà présentée. Par exemple, dans la figure (6.2f), les termes de la requête "*pakistan factory fire*" ont la même distribution temporelle sur plusieurs intervalles de temps (e.g., les 50 premières heures). Un point important qui peut être soulevé ici est quand un terme d'une requête apparaît dans beaucoup de documents pertinents et non pertinents dans la même période de temps. Dans ce cas là, nous assumons que la prise en compte de la dépendance temporelle entre ce terme et les autres termes permettrait de discriminer les documents pertinents de ceux qui ne le sont pas. Cette hypothèse de corrélation temporelle est analogique au concept de mesures proximité entre les termes (Zhu *et al.*, 2012), mais avec une proéminence sur l'aspect temporel. Dans la suite de ce chapitre, nous exploitons cette propriété pour répondre aux questions de recherche suivantes :

1. **QR1**. Dans quelle mesure les termes des requêtes sont temporellement corrélés dans les documents pertinents ?

2. **QR2**. Comment représenter et exploiter cette corrélation dans un modèle d'ordonnancement sensible au temps ?

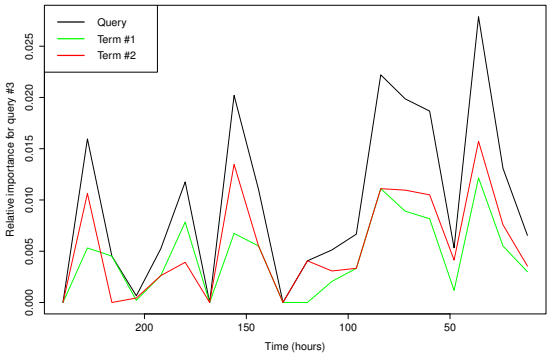
3. **QR2**. Comment estimer la pertinence temporelle des documents et la combiner avec d'autres critères de pertinence pour améliorer la RI sensible au temps ?

2. `https://www.trec-ts.org`

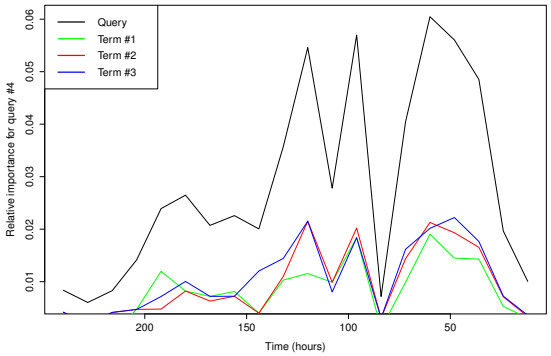


Q1 : *"buenos aires train crash"*

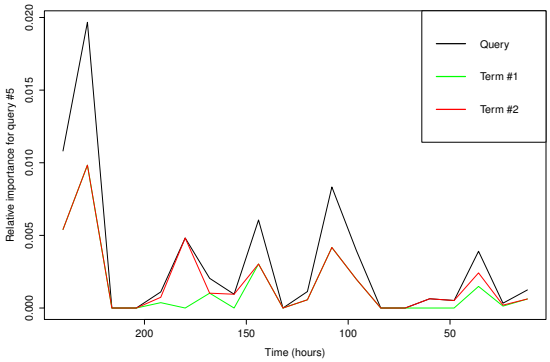
Q2 : *"pakistan factory fire"*



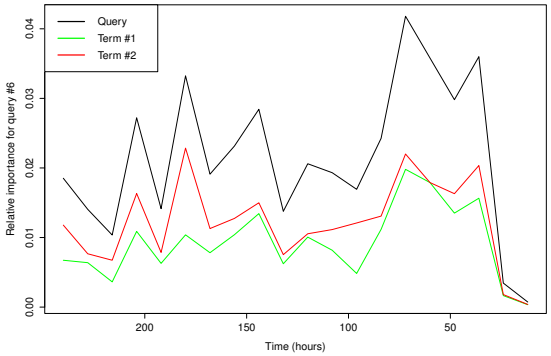
Q3 : *"colorado shooting"*



Q4 : *"sikh temple shooting"*



Q5 : *"buenos aires train crash"*



Q6 : *"pakistan factory fire"*

FIGURE 6.2: Séries temporelles des documents pertinents de 6 requêtes ($Q1$–$Q6$) de la collection utilisée par la tâche Temporal Summarization de TREC 2013.

6.3 Modèle d'agrégation sensible au temps : exploitation des dépendances temporelles entre les termes de requête

Notre approche sensible au temps est basée sur un modèle de RI (Li et Croft, 2003; Berberich *et al.*, 2010; Dakka *et al.*, 2012) qui intègre le facteur temporel dans un modèle de langue. Nous commençons tout d'abord par l'introduction du modèle de langue classique, suivi d'une présentation de ce modèle de langue temporel. Ensuite, nous décrivons comment nous avons exploité ce modèle ainsi que la corrélation temporelle des termes des requêtes pour répondre à des besoins dépendant du temps.

6.3.1 Formalisation du problème

Considérons une requête $q = w_1, w_2, \ldots, w_n$, où w_i est un terme de la requête, et un document $d_j^t \in \mathcal{D}$, où t est le temps de publication de d_j. t est le nombre de secondes écoulées depuis le $1er$ janvier $197000:00:00$ UTC jusqu'à la date de publication du document.

L'objectif principal d'un système de RI sensible au temps consiste à retourner les documents pertinents qui sont publiés dans des intervalles de temps qui sont d'intérêt pour la requête q. Notre contribution majeure ici consiste à exploiter les indicateurs temporels qu'on peut extraire à partir de la requête (également à partir de ses termes) pour l'intégrer au cœur du modèle. Comme le calcul pertinence concerne principalement les requêtes, le premier défi consiste à trouver la méthode appropriée pour combiner le facteur temporel et d'autres facteurs de pertinence dans un seul modèle d'ordonnancement pour tous les termes de la requête. Dans ce travail, nous nous limitons uniquement au facteur de pertinence thématique en plus du facteur temporel.

6.3.2 Modèle

Notre modèle inclut deux étapes principales, illustrées dans l'algorithme 2 :

– La première étape consiste à calculer la pertinence des documents suivant chaque terme de la requête, suivant le critère temporel $P(t|w_i)$ et le critère thématique $P(w_i|d_j)$. Cette étape donne lieu à un nombre de listes

d'ordonnancements associées à chaque terme. Ensuite, nous identifions les *laps* de temps des top K documents les mieux classés dans chaque liste. Ainsi, nous définissons un ensemble de périodes de temps saillantes de cet ensemble de documents. Une période de temps saillante est estimée par la moyenne des estampilles des top K documents retournés selon chaque terme de la requête.

– Dans la deuxième étape, nous fusionnons les listes des ordonnancements obtenues selon les termes des requêtes en une seule liste résultante. L'objectif de cette étape est de *booster* les documents qui sont publiés dans les mêmes périodes de temps qu'un nombre important de documents pertinents retournés en réponse à tous les termes de la requête.

Le tableau 6.1 décrit les notations utilisées dans l'algorithme 2.

Notation	**Description**
n	Nombre de termes de la requête
q	Une requête contenant n termes $(w_1, w_2, \ldots, w_n)$
d^t	Un document publié au temps t
r_{w_i}	Une liste d'ordonnancement retournée en réponse du terme w_i, $r_{w_i} \in R$
R	L'ensemble des listes d'ordonnancements en réponse des termes de la requête

TABLE 6.1: L'ensemble des notations utilisées dans l'algorithme 2.

Dans ce qui suit, nous détaillons les deux étapes de notre approche.

6.3.2.1 Génération des ordonnancements des termes de requêtes

Notre approche se base en grande partie sur le modèle de langue sensible au temps proposé dans (Dakka *et al.*, 2012), basé à son tour sur le modèle probabiliste sensible au temps de Li et Croft (2003). Comme nous l'avons déjà mentionné, nous commençons par appliquer le modèle au niveau de la requête (i.e., chaque terme est perçu comme une requête à part). Le modèle proposé ordonne les documents dans l'ordre décroissant de leur probabilité de pertinence en se basant sur les critères temporel ($P(t|w_i)$) et thématique ($P(t|w_i)$). Ainsi, le score global des documents est calculé comme suivant :

Algorithm 2: Les différentes étapes du modèle.

Entrées: $q = w_1, w_2, \ldots, w_n$, n, $d_j^t \in \mathcal{D}$.
Sortie: Ordonnancement des documents

Étape 1 : calcul de la pertinence suivant chaque terme de la requête

1. **Pour** $i = 1$ à n {*Chaque terme est considéré comme une requête à part entière*} **Faire**
2. Calculer le facteur de correspondance thématique (topicalité) d'un document $P(w_i|d_j^t)$ $(j \in 1 \ldots |\mathcal{D}|)$
3. Calculer le facteur temporel $(P(t|w_i))$
4. Calculer le score global de chaque document (combinaison des deux dimensions topicale et temporelle)
5. r_{w_i} := Top K documents retournés suivant w_i
6. **Fin Pour**

Step 2 : Fusion d'ordonnancements

7. Identifier la moyenne des estampilles de r_{w_i} $(i : 1 \ldots n)$
8. Fusionner la liste ordonnée de chaque terme de requête
9. **Retourner** la liste d'ordonnancement suivant la requête q

$$P(d^t|w_i) = P(d, t|w_i) \propto P(d|w_i)P(t|w_i) \tag{6.1}$$

$$\propto P(q|w_i)P(d)P(t|w_i) \propto P(w_i|d)P(t|w_i) \tag{6.2}$$

Où $P(w_i|d)$ dénote le degré de vraisemblance du terme w_i dans le document d, et $P(d)$ est la probabilité a priori que d pertinent pour terme de requête. Étant donné que $P(d)$ est uniforme, elle peut être ignorée dans la formule de calcul. Vu que w_i est un terme lexical, $P(w_i|d)$ peut être estimée en utilisant un modèle de vraisemblance standard. Pour éviter le problème de la probabilité nulle (i.e., probabilité que le terme soit généré du document), nous utilisons le lissage Dirichlet (Mackay et Peto, 1995) donné comme suit :

$$P(w_i|d) = \frac{tf(w_i, d) + \mu.\frac{tf(w_i,d)}{|D|}}{|d| + \mu} \tag{6.3}$$

Où $tf(w_i, d)$ est la fréquence de w_i dans d.

Le facteur $P(t|w_i)$ représente l'importance relative de t pour le terme w_i. Cette pertinence temporelle peut être estimée en utilisant différentes mé-

thodes, comme par exemple, le *maximum de vraisemblance*, défini comme la somme normalisée des scores de pertinence des documents publiés à l'instant t pour le terme w_i :

$$P(t|w_i) = \frac{tf(w_i, D^t)}{|D^t|} \tag{6.4}$$

Où D^t est l'ensemble des documents publiés à l'instant t. Il est à noter que cette fonction de pondération assume une indépendance temporelle entre les termes de la requête. Les listes d'ordonnancements obtenues dans cette phase donnent résultat à différentes listes $r_w \in R$ selon les deux dimensions de pertinence thématique et temporelle. L'agrégation de ces listes est détaillée dans la section qui suit.

6.3.2.2 Agrégation d'ordonnancements sensible au temps

Pour fusionner les listes d'ordonnancements obtenues dans la première étape de notre approche, nous adaptons la méthode de fusion d'ordonnancements *Reciprocal Rank Fusion* (Cormack *et al.*, 2009), en injectant une distance de proximité temporelle qui tire profit de la dépendance temporelle entre les termes de la requête. Pour la concrétiser, nous appliquons une variante normalisée de la fonction gaussienne d'estimation de densité. Les scores des documents donnés par notre modèle d'ordonnancement sensible au temps TTD-M (*Temporal Term Dependent Model*) sont calculés comme suit :

$$TTDM(d^t \in D) = \sum_{r \in R} \frac{1}{\epsilon + r(d_t)} * kernel(t, t_{avg}) \tag{6.5}$$

Où $r_w(d^t)$ est la position du document d dans la liste de résultat r_w, et t_{avg} est la moyenne des *estampilles* des documents les mieux classés dans R. Nous assumons que t_{avg} est la période de temps la plus importante pour une requête donnée. Cette hypothèse favorise les documents retournés par tous (ou la plupart) des termes de la requête, qui sont publiés dans des périodes de temps très proches des top K documents les mieux classés. Ceci dit, si deux documents sont suffisamment proches en terme de pertinence thématique et temporelle, pour tous les termes de requête, alors ils doivent être bien classés dans la liste finale des résultats.

La fonction de densité gaussienne est donnée par :

$$kernel(t_1, t_2) = \frac{1}{\sqrt{2\pi\sigma}} * exp[\frac{-(t_1 - t_2)^2}{2\sigma^2}] \tag{6.6}$$

où σ désigne la variance de la densité du noyau.

6.4 Évaluation expérimentale

Dans cette section, nous présentons le cadre expérimental et nous décrivons la collection de données ainsi que les référentiels de comparaison et les mesures d'évaluation utilisées. Ensuite, nous présentons les résultats de l'analyse de corrélation temporelle des termes des requête, ainsi qu'une analyse comparative des performances avec des modèles de RI temporels standards et des méthodes de RI non sensibles au temps.

6.4.1 Cadre expérimental

6.4.1.1 Données expérimentales et tâche de recherche

Nous utilisons la collection de données Stream Corpus [3] fournie par la tâche KBA de TREC et exploitée par la tâche TS [4] de TREC 2013 et 2014. Le corpus comporte plus de 500 millions de documents issus de plusieurs sources (Presse, Web, Social, Forum, Blog, etc.) avec une taille de 4.5 Téra octets compressés. Tous les documents sont datés dans la période allant du mois d'octobre 2011 jusqu'au mois de février 2013. La collection a été indexée en utilisant *Lucene* [5]. Chaque document contient un ensemble de phrases avec un identifiant unique. La tâche propose 10 topics en 2013 et 14 topics en 2014, où chaque topic fait référence à un événement du monde réel. Ces topics correspondent à des événements d'actualité tels que des manifestations, des accidents ou des catastrophes naturelles. La figure 6.3 présente un exemple de requête ($Q1$) qui correspond à l'événement *buenos aires train crash*.

3. `http://trec-kba.org/kba-stream-corpus-2013.shtml`
4. `http://www.trec-ts.org/`
5. `https://lucene.apache.org/`

```
<event>
    <id>1</id>
    <title>2012 Buenos Aires Rail Disaster</title>
    <description>http://en.wikipedia.org/wiki/2012_Buenos_
    Aires_rail_disaster</description>
    <start>1329910380</start>
    <end>1330774380</end>
    <query>buenos aires train crash</query>
    <type>accident</type>
</event>
```

FIGURE 6.3: Requête $Q1$ "*buenos aires train crash*" de la tâche TS 2013.

L'objectif de cette tâche est de concevoir des systèmes permettant de surveiller les événements en détectant à la volée toutes les informations "nouvelles" publiées en temps réel. Cette tâche présente clairement un caractère temporel flagrant vu que les systèmes participants doivent être en mesure de retourner des résultats qui sont à la fois temporellement et thématiquement pertinents.

***Id* de la Requête**	**Requête**	**#Documents pertinents**
1	*Q1 : Buenos Aires Crash Train*	789
2	*Q2 : Factory Fire Pakistan*	585
3	*Q3 : Colorado Shooting*	243
4	*Q4 : Shooting Sikh Temple*	613
5	*Q5 : Hurricane Isaac*	36
6	*Q6 : Hurricane Sandy*	518
7	*Q7 : Derecho midwest*	2
8	*Q8 : Bopha Typhoon*	210
9	*Q9 : Earthquake Guatemala*	294
10	*Q10 : Tel Aviv Bombing Bus*	284

TABLE 6.2: Requêtes de la tâche TREC TS 2013 avec les documents pertinents associés.

Le tableau 6.2 présente également quelques statistiques sur les requêtes et les documents pertinents associés à la tâche TS de TREC 2013. Comme nous le voyons à partir du tableau, les requêtes ne disposent pas du même nombre de documents pertinents. Certaines requêtes ont un nombre très faible de

documents (eg., $Q5$ et $Q7$), ce qui rend la tâche de recherche un peu difficile pour le modèle se basant sur la dimension thématique. Le nombre de termes des requêtes varie entre 2 et 4, ce qui correspond généralement à la moyenne dans les systèmes de RI réels (Baeza-Yates et Ribeiro-Neto, 1999).

6.4.1.2 Mesures d'évaluation

Nous évaluons notre approche ainsi que les référentiels de comparaison en utilisant les mesures de rappel, précision et F-mesure, sur lesquelles se sont basées les organisateurs de la tâche TREC TS pour évaluer les systèmes des groupes participants :

$$Précision = \frac{|Documents\ pertinents\ retournés|}{|Documents\ retournés|} \tag{6.7}$$

$$Rappel = \frac{|Documents\ pertinents\ retournés|}{|Documents\ pertinents|} \tag{6.8}$$

$$F\text{-}mesure = 2 * \frac{Précision * Rappel}{Précision + Rappel} \tag{6.9}$$

Il est à noter que nous avons effectué l'évaluation au niveau des documents, et non pas au niveau des phrases tel est le cas avec la tâche officielle. En effet, notre objectif est d'étudier de la capacité de notre modèle à retourner des documents (temporellement) pertinents en réponse à des requêtes sensibles au temps. L'évaluation au niveau des phrases consiste à vérifier, en plus de la pertinence des documents, la pertinence des phrases qui les constituent. Nous avons conduit l'analyse de corrélation sur l'ensemble des requêtes et documents de la tâche TS 2013, et nous avons évalué les performances de recherche de notre modèle et les référentiels de comparaison en utilisant les données fournies par la tâche TS 2014.

6.4.1.3 Référentiels de comparaison

Nous comparons notre modèle d'ordonnancement avec les méthodes suivantes :

- Un modèle atemporel : le modèle de langue (ML) avec un lissage Dirichlet (Zhai et Lafferty, 2004) (comme présenté dans l'Eq. 6.3) ;
- Deux modèles sensibles au temps :

– Le modèle de langue temporel proposé par Dakka *et al.* (2012) (comme présenté dans l'Eq. 6.1) ;
– Le modèle *Recency Prior* (RP) proposé par Li et Croft (2003). Ce dernier définit une distribution à priori sur les documents afin de favoriser les documents les plus récents :

$$P(d) = \lambda e^{-\lambda t_d} \tag{6.10}$$

Où λ est le taux d'une distribution exponentielle et t_d et l'age du document d.

6.4.2 Analyse de corrélation temporelle entre les termes des requêtes

Afin de valider notre intuition sur la corrélation temporelle, conformément à notre première question de recherche (QR1), nous commençons d'abord par examiner à quel point les termes des requêtes sont dépendants. La figure 6.4 montre la matrice de similarité illustrant la corrélation entre les séries temporelles (Ng *et al.*, 2001) des termes des requêtes de la tâche TS 2013. Cette corrélation temporelle mesure la similarité entre deux séries chronologiques en se basant sur l'importance relative de chaque terme (Cf., Eq. 6.4).

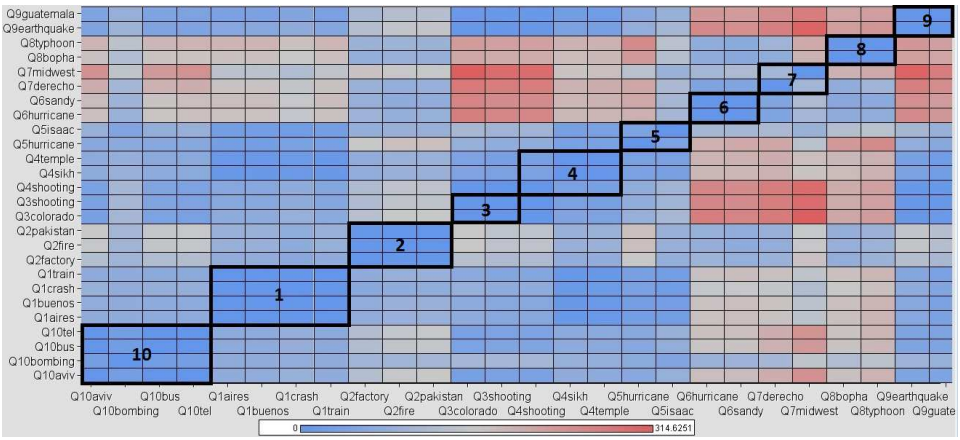


FIGURE 6.4: Analyse de corrélation temporelle des termes de requêtes de la tâche TREC TS 2013.

Les colonnes et les lignes de la matrice indiquent les termes appartenant aux requêtes ($Q1 - Q10$), donc 26 termes en total. Les termes des requêtes

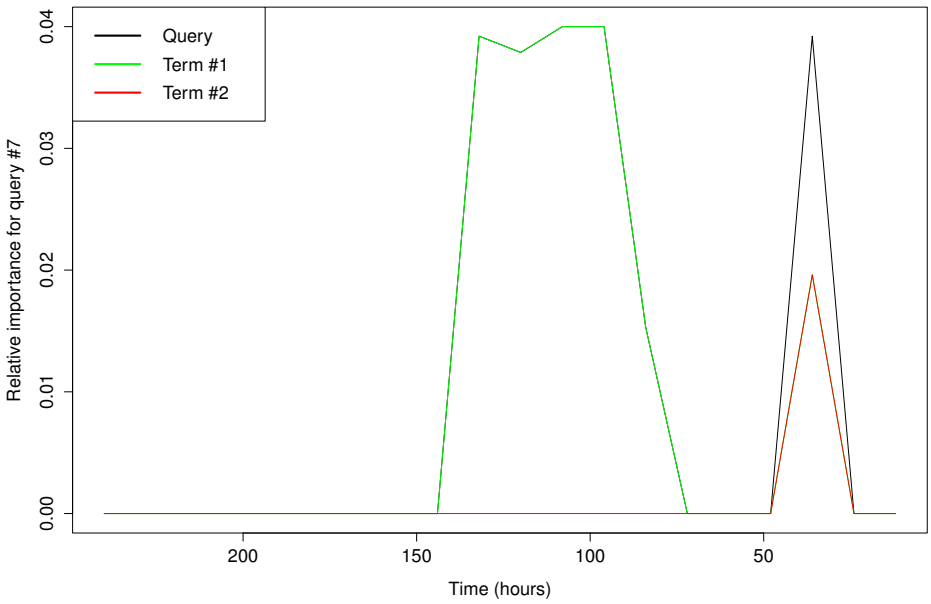


FIGURE 6.5: Distribution des termes de la requête Q7 ("*midwest derecho*") au cours du temps, dans les documents pertinents de la tâche TS 2013. L'axe des abscisses représente le temps en heure, et l'axe des ordonnées représente poids normalisé de la requête et ses termes dans les documents.

sont ordonnés suivant l'*Id* de la requête sur les deux axes. Les blocs en diagonale encadrés par des rectangles en gras sur la matrice correspondent aux mesures de similarité entre les termes d'une même requête. Plus la couleur est d'un niveau de bleu accentué, plus la similarité est grande. La couleur rouge indique un degré de similarité faible. Nous pouvons constater à partir de la figure 6.4 (en se basant sur l'intensité des couleurs) que les séries chronologiques des termes appartenant à la même requête sont plus corrélées que celles ne faisant pas partie de la même requête (i.e., les cases en dehors des rectangles en gras). Cette observation concorde avec notre hypothèse sur la dépendance temporelle entre les termes d'une même requête. Nous pouvons aussi remarquer que les termes "*midwest*" (*terme*#1) et "*derecho*" (*terme*#2) de la requête *Q*7 ne sont pas temporellement dépendants. Cette observation peut être expliquée de deux façons. La première raison est que le nombre de documents pertinents selon cette requête est assez faible, comme indiqué dans le tableau 6.2 (Cf., section 6.4.1.1), ce qui entraîne l'absence des

termes dans la collection. Une deuxième raison possible est que les termes de cette requête ne sont pas assez discriminants dans la collection, c'est à dire qu'ils existent dans de nombreux autres documents.

Nous poursuivons cette analyse en montrant sur la figure 6.5 la distribution temporelle des termes appartenant à la requête $Q7$ "*midwest derecho*" dans le temps, dans les documents pertinents de la tâche TS 2013. La figure montre que le terme "*midwest*" ($terme\#1$) apparaît dans une grande partie des documents pertinents entre les premières 60 heures et 150 heures après l'occurrence de $Q7$. En effet, "*midwest*" n'est pas un terme clé qui pourrait décrire cet événement, car il est aussi fréquent dans des documents pertinents faisant référence à d'autres requêtes.

6.4.3 Résultats expérimentaux

Dans cette section, nous présentons les résultats obtenus par notre modèle sur les requêtes de la collection TS 2014. Nous commençons tout d'abord par paramétrer notre modèle ainsi que les référentiels de comparaison. Le paramètre de lissage μ du modèle de langue (ML) est fixé à 2000, tandis que le paramètre λ du modèle *Rencency Prior* (RP) est fixé à 0.01 (comme dans Efron et Golovchinsky (2011)). Les paramètres ϵ de la méthode de fusion RRF et σ de la fonction de densité gaussienne sont empiriquement fixés à 30 et 170, respectivement. Dans chaque expérimentation, nous utilisons le modèle de lissage Dirichlet pour extraire les 100 documents de chaque heure de notre collection, en réponse à chacune des requêtes. Ensuite, nous appliquons notre modèle pour les ré-ordonnancer.

	Précision	**Rappel**	**F-Mesure**	**% ↗**
ML	0.0830	0.2019	0.1177	+32.47% ⋆
MLT	0.1307	0.1772	0.1504	+13.71% ⋆
RP	0.0866	0.2019	0.1212	+30.46%
TTD-M	**0.1692**	**0.1797**	**0.1743**	-

TABLE 6.3: Analyse comparative des performances de notre modèle d'ordonnancement sensible au temps (TTD-M). % ↗ indique le taux d'accroissement en terme de F-mesure, et le symbole "⋆" dénote le test de significativité : "⋆" : $t < 0.05$.

Le tableau 6.3 montre les résultats obtenus en terme de précision, rappel et F-mesure, ainsi que les taux d'accroissement et les test de significativité de *student*. Nous pouvons observer à partir du tableau 6.3, une amélioration de +32.74% en terme de F-mesure qui est mise en avant par notre modèle TTD-M par rapport aux modèles de référence.

Id	**Termes de la requête**	**F-Mesure**		
		TTD-M	**RP**	% ↗
11	*costa concordia*	0,2055	0,0904	**+55,98**%
12	*european cold wave*	0,0763	0,0347	**+54,49**%
13	*queensland floods*	0,2262	0,0787	**+65,21**%
14	*boston marathon bombing*	0,0802	0,1171	-45,99%
15	*egyptian riots*	0,1525	0,1028	**+32,56**%
16	*quran burning protests*	0,3646	0,2352	**+35,47**%
17	*in amenas hostage crisis*	0,1252	0,2361	-88,59%
18	*russian protests*	0,2107	0,0971	**+53,89**%
19	*romanian protests*	0,3470	0,0794	**+77,10**%
20	*egyptian protests*	0,0831	0,0727	**+12,48**%
21	*russia meteor*	0,0707	0,143	-100%
22	*bulgarian protests*	0,1967	0,0606	**+69,15**%
23	*shahbag protests*	0,0281	0,0489	-73,92%
24	*nor'easter*	0	0	0
25	*Southern California shooting*	0,0057	0,0510	-100%

TABLE 6.4: Analyse au niveau des requêtes des performances de notre modèle (TTD-M) vs. le modèle RP. La dernière colonne (% ↗) indique le taux d'accroissement en terme de la métrique F-Mesure.

Ce résultat est prévisible étant donné que le cadre de RI dépend fortement de la dimension temporelle et que le modèle de langue est basé uniquement sur le critère thématique. D'autre part, notre modèle fournit des améliorations un peu moins significatives en terme de F-mesure autour de +13.71% par rapport au modèle de langue temporel MLT. Comme présenté dans l'Eq.6.1, le score temporel favorise les documents qui sont publiés dans des intervalles de temps qui peuvent être intéressants pour la requête. De plus, les performances de notre approche par rapport à la méthode RP sont largement meilleurs, allant jusqu'à +30.46%. Cependant cette différence n'est pas significative. Ceci peut être expliqué par le fait que les documents pertinents

sont généralement distribués d'une manière uniforme sur des périodes de temps différentes. Ceci dit, il existe des requêtes pour lesquelles la fraîcheur n'est pas un facteur très important. Par conséquent, retourner uniquement les documents récents peut dégrader la précision du modèle.

Nous poursuivons ces analyses en présentant dans le tableau 6.4 une analyse au niveau des requêtes, des performances de notre modèle TTD-M en comparaison avec le modèle RP pour chaque requête de la tâche TS 2014. Le tableau 6.4 montre que TTD-M fournit des meilleurs résultats que RP pour 60% des requêtes (9/15). Une analyse des requêtes avec des taux d'accroissement positifs montre que la plupart de leurs documents pertinents associés sont distribués d'une manière uniforme sur des périodes de temps différentes avec des pics sur des intervalles spécifiques. La figure 6.6 présentant la distribution des documents pertinents de la requête $Q11$ (cf., 6.4) illustre cette observation.

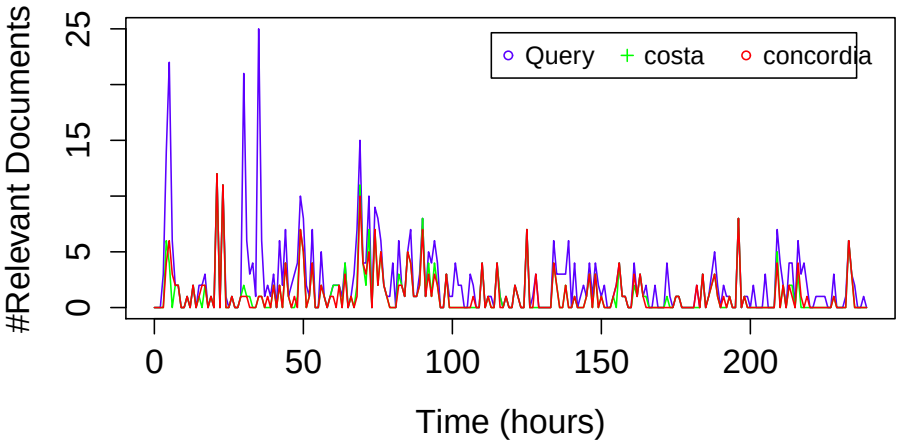


FIGURE 6.6: Série chronologique des termes de la requête $Q11$ ("*costa concordia*") dans les documents pertinents de la tâche TS 2014.

Contrairement au modèle RP, TTD-M exploite le facteur temporel comme pour le modèle MLT, et étend ce dernier en favorisant les documents présentant une importance relative similaire dans le temps de tous les termes de la requête, comme montré dans la figure 6.6. Nous pouvons conclure que

l'utilisation de RRF conjointement avec la fonction de densité gaussienne permet d'améliorer l'ordonnancement des documents populaires au sein des différentes listes qui sont retournés suivant les termes d'une même requête. En effet, tous les documents retournés dans ces différentes listes, dans des fenêtres temporelles similaires sont bien classés dans la liste d'ordonnancement finale. Pour confirmer cette corrélation temporelle, qui est notre hypothèse initiale, nous montrons à travers la figure 6.7 la matrice de corrélation de la requête $Q11$ ainsi que les termes qui la constituent.

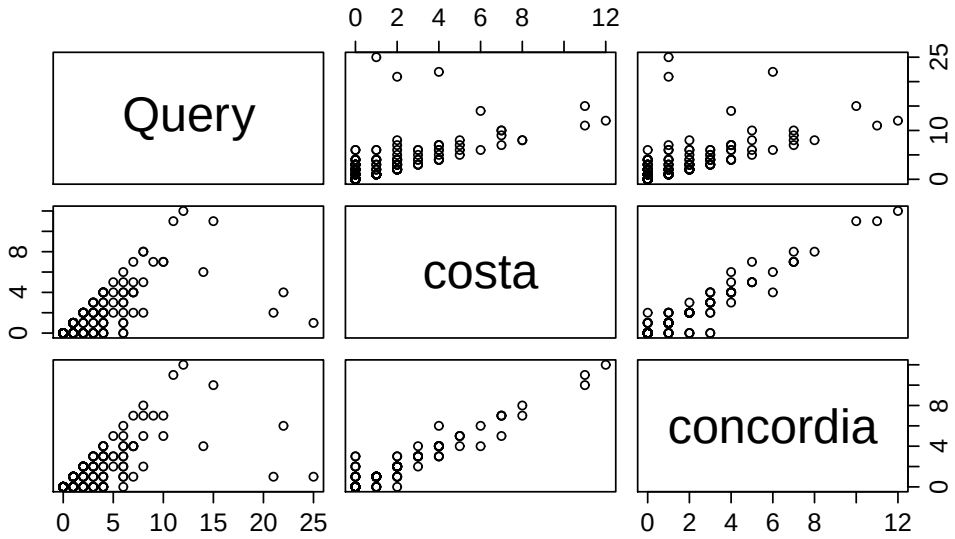


FIGURE 6.7: Corrélation entre les termes de la requête $Q11$ ("*costa*" et "*concordia*") avec la même requête Q11.

La figure 6.7 montre encore que les deux termes "*costa*" et "*concordia*" sont temporellement corrélés avec une valeur de corrélation autour de 0.96. La corrélation des deux termes avec la requête entière est aussi significative avec des valeurs de 0.6725 et 0.6702 pour "*costa*" et "*concordia*", respectivement. Il est à noter aussi que le modèle RP améliore notre modèle pour certaines requêtes (5/15). Par exemple, pour la requête $Q21$ ("*russia meteor*"), la différence est d'environ 100%. À partir de l'analyse de la distribution temporelle des documents pertinents pour cette requête, présentée dans la figure 6.8, nous pouvons justifier ces performances par deux faits. D'abord, quand

on utilise uniquement des termes individuels comme des requêtes, on peut avoir un très grand nombre de documents qui sont uniformément distribués sur une période de temps relativement longue. Ainsi, quand nous appliquons la méthode de fusion conjointement la fonction de densité, il est difficile de garder que les documents distribués sur une période de temps plus courte, i.e., les 50 premières heures comme montré dans la figure 6.8.

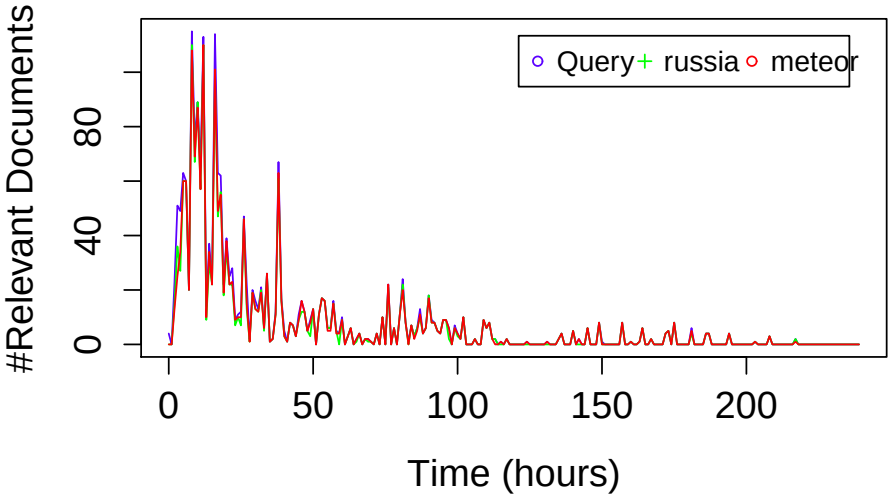


FIGURE 6.8: Les séries chronologiques de la requête $Q21$ et ses termes dans les documents pertinents de la tâche TREC TS 2014.

Par ailleurs, bien que les valeurs de F-mesure semblent être relativement faibles, ceci semble être raisonnable, étant donné que le modèle RP qui est uniquement performant quand il s'agit des requêtes s'intéressant à des documents récents. Par contre, ce modèle n'est pas adapté pour répondre à des requêtes comme ($Q21$).

6.5 Conclusion

Dans ce chapitre, nous avons proposé un modèle d'ordonnancement sensible au temps permettant d'injecter la dimension de pertinence temporelle

dans une architecture d'agrégation d'ordonnancements. Plus particulièrement, nous nous basons sur une distance de proximité temporelle entre les termes d'une même requête pour retourner des documents qui satisfont à la fois le critère thématique et temporel. Ainsi, nous nous sommes basés sur une méthode de fusion d'ordonnancements pour formuler la tâche comme un problème de fusion de données, où chaque terme de requête est considéré comme une requête en soi. Nous avons mené une évaluation expérimentale en utilisant une large collection de données standard qui est fournie par les tâches *Knowledge Base Acceleration* (KBA) et *Temporal Summarization* (TS) de TREC 2013 et 2014. Les analyses expérimentales ont montré l'impact positif de la prise en compte des corrélations temporelles entre les termes des requêtes dans notre modèle sur les résultats de recherche. L'exploitation de l'aspect temporel dans la méthode d'agrégation d'ordonnancements a également permis de favoriser les documents qui sont publiés dans des périodes de temps pertinentes pour les requêtes.

Troisième partie

Conclusion générale

Chapitre 7

Conclusion générale

7.1 Synthèse des contributions

Les travaux présentés dans ce manuscrit s'inscrivent dans le contexte de l'agrégation multicritères, qui correspond à un des domaines émergents de la RI avec de nombreux enjeux, tels que formalisation de modèles d'ordonnancement, la prise en compte des différents critères de pertinence impactant les jugements de pertinence des utilisateurs et la fusion de données, etc.

Dans cette thèse, nous nous sommes particulièrement intéressés à la proposition et l'évaluation de modèles d'agrégation de pertinence multidimensionnelle en RI. Ainsi, nous avons axé notre revue de l'état de l'art selon cette dimension en donnant un aperçu des différentes approches qui ont été proposées. Ces approches ont été fondées principalement sur les modèles d'agrégation classiques et prioritaires, les méthodes de surclassement et les approches d'agrégation et d'apprentissage d'ordonnancements. Une première catégorie de travaux reposait principalement sur des combinaisons linéaires simples pour l'agrégation des différentes dimensions de pertinence. Cependant, ces travaux se basent sur l'hypothèse non réaliste d'additivité ou d'indépendance des dimensions, ce qui rend le modèle non approprié dans plusieurs situations réelles dans lesquelles les critères étant corrélés ou présentant des interactions entre eux. Pour répondre à cet enjeu, une deuxième famille d'approches propose d'appliquer des techniques issues du domaine de l'apprentissage automatique, permettant ainsi d'apprendre un modèle par

l'exemple et de le généraliser dans l'ordonnancement et l'agrégation des critères. Toutefois, ces méthodes d'apprentissage d'ordonnancements ont tendance à offrir un aperçu limité sur la façon de considérer l'importance et l'interaction entre les critères qui représentent les différentes dimensions de pertinence. Outre la sensibilité des paramètres utilisés dans ces algorithmes, il a été très difficile de comprendre pourquoi un critère est préféré par rapport à un autre.

Dans cette direction de recherche, nous avons proposé deux types de contributions. Tandis que la première est basée sur une méthode d'agrégation multicritères issue du domaine de prise de décision floue, la deuxième s'appuie sur une approche d'agrégation d'ordonnancements sensible au temps s'appliquant dans des flux de documents qui changent au cours du temps. Nous les décrivons brièvement ci-dessous :

1. Un modèle de combinaison de pertinence multicritères basé sur un opérateur d'agrégation flou. Ce dernier permet la modélisation des interactions entre les critères, et ce à travers une mesure floue définie sur l'ensemble des dimensions de pertinence. Cette mesure permet de surmonter le problème d'additivité des fonctions de combinaison classiques, qui sont incapables de modéliser plusieurs situations du monde réel. Ainsi, nous avons adapté ce modèle pour deux scénarios de combinaison de pertinence multicritères : *(i)* un cadre de recherche d'information multicritères dans un contexte de recherche de tweets ; et *(ii)* deux cadres de recherche d'information personnalisée pour tester l'efficacité du modèle à adapter ses résultats suivant les préférences des utilisateurs. Pour chacun de ces deux modèles, nous avons proposé un algorithme d'apprentissage des degrés d'importance des critères, ce qui permet ainsi de donner une idée plus claire sur les corrélations et interactions entre les critères.

2. Intégration de la dimension de pertinence temporelle dans le processus d'ordonnancement des documents. Tout d'abord, nous avons souligné l'importance du facteur temporel surtout dans les flux de documents qui changent dans le temps. Ainsi, nous avons proposé un modèle d'agrégation sensible au temps pour le combiner avec le facteur de pertinence thématique. Dans cet objectif, nous avons effectué une analyse temporelle pour éliciter l'aspect temporel des requêtes, et nous avons proposé une évaluation de ce modèle dans une tâche de recherche sensible au temps.

7.2 Perspectives

Les différentes évaluations expérimentales menées pour évaluer nos différentes contributions ont montré leur efficacité vis-à-vis d'un ensemble de modèles représentatifs de l'état de l'art. Ce manuscrit ouvre de nombreuses perspectives que nous synthétisons dans ce qui suit.

À court terme, nous souhaitons améliorer nos contributions selon trois aspects :

1. *Apprentissage de l'importance des critères de pertinence.* Nous avons proposé un algorithme d'apprentissage générique, et nous l'avons appliqué uniquement dans des cadre de recherche d'information impliquant au maximum trois dimensions de pertinence. Une des limites de cet algorithme est sa complexité quand il s'agit d'apprendre l'importance de plus de trois critères. Nous notons toutefois que cette complexité est indépendante du processus d'agrégation, et elle est uniquement liée à l'apprentissage des paramètres car elle s'effectue "hors ligne", contrairement aux algorithmes d'apprentissage automatique où elle est liée à la fois aux deux phases d'apprentissage et de test (Liu, 2009). Nous pensons qu'il pourrait être intéressant de lever ce verrou en allégeant les contraintes d'additivité sur l'ensemble des critères, ce qui permet de réduire la complexité de l'algorithme d'apprentissage des capacités.

   Un deuxième problème qui est aussi lié à cet algorithme, malgré sa capacité d'interpréter l'importance et la corrélation entre les critères, est le fait qu'il est contraint par la présence d'un ensemble d'apprentissage. Ainsi, nous proposons comme perspective de recherche de généraliser le modèle pour faire face aux problèmes d'agrégation où l'on dispose pas de données préalables sur les critères et les documents (tels que des scores partiels et globaux).

2. *Agrégation en l'absence de scores.* Nos deux modèles d'agrégation sont basés sur l'hypothèse de présence des scores de pertinence suivant tous les critères considérés. Toutefois, nous pourrons être confrontés à des situations de RI où l'on dispose uniquement des positions (*ranks*) des documents. Cet axe de recherche est principalement traité par les méthodes d'agrégation d'ordonnancements (Aslam et Montague, 2001), et plus spécifiquement par les méthodes positionnelles (Renda et Straccia, 2003). Il serait intéressant d'intégrer les deux familles de méthodes d'agrégation dans un seul cadre de combinaison multicritères qui peut s'adapter à toutes les situations de recherche possibles.

A moyen terme, nous proposons de généraliser nos contributions selon deux dimensions, liées aux scénarios de recherche sensibles au temps :

1. *Généralisation aux différentes tâches.* Dans notre modèle de RI temporel, nous nous sommes intéressés à l'intégration du temps dans le processus d'ordonnancement des documents. Comme nous l'avons déjà montré dans l'évaluation expérimentale, ce modèle est plus performant pour les requêtes sensibles au temps, indépendamment de la période de temps à laquelle se réfère chaque requête. Par contre, ce dernier n'est pas efficace pour les requêtes présentant plusieurs périodes de temps saillantes, vu la méthode *basique* (i.e., une moyenne) que nous avons utilisée pour sélectionner ces périodes. L'exploitation d'une méthode d'identification plus appropriée des périodes *rafales* pour une requête spécifique pourrait donner des résultats plus précis (Kleinberg, 2002; Zhu et Shasha, 2003).

   Nous notons aussi que les caractéristiques temporelles que présentent les requêtes au cours du temps peuvent révéler des différents types de requêtes, comme déjà discuté dans le chapitre 4. Une perspective intéressante consiste à généraliser notre modèle pour qu'il soit capable d'éliciter automatiquement le caractère temporel des requêtes et guider la phase d'ordonnancement en fonction du type de la requête. Ce traitement pourrait être également généralisé dans d'autres tâches de RI autres que celles proposées par TREC. La sous tâche *Temporal Information Retrieval* (TIR) proposée dans le cadre de la tâche *Temporalia* de *NTCIR* correspond parfaitement à cette perspective (Joho *et al.*, 2014). Il est aussi à signaler dans ce même contexte, que le modèle ainsi que la tâche considérée dans notre cadre de RI s'appuient sur une collection de documents dont tous les documents possèdent des *estampilles* (e.g., date de création). Toutefois, dans plusieurs situations de recherche du monde réel, ces dates peuvent ne pas être disponibles. Une des solutions possibles est de se baser sur les techniques d'extraction d'expressions temporelles pour estimer les dates auxquelles fait référence un document puis en tenir compte lors de l'estimation de pertinence (Strötgen *et al.*, 2012).
2. *Présentation des résultats de recherche.* L'objectif d'une méthode de RI est de satisfaire le besoin en information d'un utilisateur. Quand ce besoin devient dépendant du temps, nous assumons que le problème doit aller plus loin que le simple ordonnancement des documents et concerne donc la présentation des résultats de recherche (Joho et Jose, 2008; Sokolan *et al.*, 2015), qui sont également sensibles au temps.

Ainsi, cette dépendance temporelle des documents remet en question le problème de représentation traditionnelle des résultats (i.e., *snippet*, titre du document, etc) surtout pour les besoins en informations traitant des *topics* liés aux actualités. Les *timelines* (Swan et Allan, 2000) sont un exemple concret de ces applications. Par exemple, pour éviter que les utilisateurs parcourent les longs textes et annonces, *Google News* présente des résultats qui peuvent être explorés suivant le temps et qui sont aussi issus de plusieurs sources d'évidence. Ces efforts pourraient être améliorés pour inclure des résumés d'événements afin d'assister les utilisateurs recherchant des besoins en informations urgentes. Bien que la tâche INEX (Bellot *et al.*, 2014) s'intéresse aux résumés des tweets en réponse à des entités, elle n'a pas encore abordé l'aspect temporel et la présentation des tweets. Cependant, la tâche *TREC Microblog timeline generation* est une piste intéressante pour cette direction de recherche. Il existe aussi quelques tentatives pour traiter cette problématique, avec quelques systèmes qui sont ouverts [1] et d'autres qui sont encore commerciaux [2].

1. `http://www.simile-widgets.org/timeline/`
2. `http://timeglider.com/widget/index.php`

Bibliographie

Abbes, R., Pinel-Sauvagnat, K., Hernandez, N. et Boughanem, M. (2013). Irit at trec knowledge base acceleration 2013 : Cumulative citation recommendation task. *In Text REtrieval Conference (TREC)*. National Institute of Standards and Technology (NIST).

Aczel, J. (1948). On mean values. *Bulletin of the American Mathematical Society*, 54(4):392–400.

Ah-Pine, J. (2008). Data fusion in information retrieval using consensus aggregation operators. *In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, pages 662–668, Washington, DC, USA. IEEE Computer Society.

Aji, A., Wang, Y., Agichtein, E. et Gabrilovich, E. (2010). Using the past to score the present : Extending term weighting models through revision history analysis. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 629–638, New York, NY, USA. ACM.

Akritidis, L., Katsaros, D. et Bozanis, P. (2011). Effective rank aggregation for metasearching. *Journal of System Software*, 84(1):130–143.

Alonso, O., Baeza-yates, R., Strötgen, J. et Gertz, M. (2011). Temporal information retrieval : Challenges and opportunities. *In 1st Temporal Web Analytics Workshop at WWW*, CEUR Workshop Proceedings, pages 1–8.

Amati, G., Amodeo, G. et Gaibisso, C. (2012). Survival analysis for freshness in microblogging search. *In Proceedings of the 21st ACM Inter-*

national Conference on Information and Knowledge Management, CIKM '12, pages 2483–2486, New York, NY, USA. ACM.

Arrow, K. J. (1974). *Choix collectif et préférences individuelles*, volume 1 de *Perspectives de l'é conomique*. Calmann-Lévy.

Aslam, J. A. et Montague, M. (2001). Models for metasearch. *In Proceedings of the 24th annual International* ACM SIGIR *Conference on Research and Development in Information Retrieval*, pages 276–284, New York, NY, USA. ACM.

Asur, S. et Buehrer, G. (2009). Temporal analysis of web search query-click data. *In SNA-KDD*, pages 1–8, Paris, France. ACM Press.

Baeza-Yates, R. A. et Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Baeza-Yates, R. A. et Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Barry, C. L. (1994). User-defined relevance criteria : An exploratory study. *Journal of the American Society for Information Science*, 45(3):149–159.

Becker, H., Naaman, M. et Gravano, L. (2011). Beyond trending topics : Real-world event identification on twitter. *In ICWSM*. The AAAI Press.

Bellot, P., Bogers, T., Geva, S., Hall, M. A., Huurdeman, H. C., Kamps, J., Kazai, G., Koolen, M., Moriceau, V., Mothe, J., Preminger, M., SanJuan, E., Schenkel, R., Skov, M., Tannier, X. et Walsh, D. (2014). Overview of INEX 2014. *In Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF*, pages 212–228.

Ben Jabeur, L., Tamine, L. et Boughanem, M. (2010). A social model for literature access : towards a weighted social network of authors. *In Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 32–39, Paris, France.

Berardi, G., Esuli, A., Marcheggiani, D. et Sebastiani, F. (2011a). ISTITREC Microblog Track 2011 : Exploring the Use of Hashtag Segmentation and Text Quality Ranking. *In Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011)*. National Institute of Standards and Technology (NIST).

Berardi, G., Esuli, A., Marcheggiani, D. et Sebastiani, F. (2011b). ISTI@TREC microblog track 2011 : Exploring the use of hashtag segmentation and text quality ranking. *In Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011)*. National Institute of Standards and Technology (NIST).

Berberich, K., Bedathur, S., Alonso, O. et Weikum, G. (2010). A language modeling approach for temporal information needs. *In Proceedings of the 32Nd European Conference on Advances in Information Retrieval*, ECIR'2010, pages 13–25, Berlin, Heidelberg. Springer-Verlag.

Blanco, R. et Zaragoza, H. (2011). Beware of relatively large but meaningless improvements. Rapport technique, Yahoo ! Research 2011-001.

Borda, J. (1781). Mémoire sur les élections au scrutin. *Histoire de l'Académie des sciences*.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10):913–925.

Bouchon-Meunier, B. et Marsala, C. (2003). *Logique floue, principes, aide à la décision*. Hermès Science, 1 édition.

Boughanem, M., Loiseau, Y. et Prade, H. (2006). Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. *In Proceedings of the Third international conference on Adaptive Multimedia Retrieval : user, context, and feedback*, AMR'05, pages 44–54, Berlin, Heidelberg. Springer-Verlag.

Boughanem, M. et Savoy, J. (2008). *Recherche d'information : état des lieux et perspectives*. Collection Recherche d'information et web. Hermès science publ. Lavoisier, Paris.

Bouidghaghen, O., Tamine, L. et Boughanem, M. (2011a). Personalizing mobile web search for location sensitive queries. *In International Conference on Mobile Data Management*, pages 110–118. IEEE Computer Society.

Bouidghaghen, O., Tamine, L., Pasi, G., Cabanac, G., Boughanem, M. et da Costa Pereira, C. (2011b). Prioritized aggregation of multiple context dimensions in mobile IR. *In Proceedings of the 7th Asia conference on Information Retrieval Technology*, volume 7097 de *AIRS'11*, pages 169–180, Berlin, Heidelberg. Springer.

Bouvier, V. et Bellot, P. (2014). Use of time-aware language model in entity driven filtering system. *In Text REtrieval Conference (TREC)*. National Institute of Standards and Technology (NIST).

Bouyssou, D., Dubois, D., Pirlot, M. et Prade, H. (2006). *Concepts et méthodes pour l'aide à la décision*, volume 1. Hermès.

Brans, J., Mareschal, B. et Vincke, P. (1984). Promethee : a new family of outranking methods in multicriteria analysis. *In* Brans, J., éditeur : *Operational Research, IFORS 84*, pages 477–490. North Holland, Amsterdam.

Brans, J. et Vincke, P. (1985). A preference ranking organization method. *Management Science*, 31(6):647–656.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Buckley, C. et Voorhees, E. M. (2000). Evaluating evaluation measure stability. *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 33–40, New York, NY, USA. ACM.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. et Hullender, G. (2005). Learning to rank using gradient descent. *In Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 89–96, New York, NY, USA. ACM.

Campos, R., Dias, G., Jorge, A. M. et Jatowt, A. (2014a). Survey of temporal information retrieval and related applications. *ACM Comput. Surv.*, 47(2):15 :1–15 :41.

Campos, R., Dias, G., Jorge, A. M. et Nunes, C. (2014b). Gte-rank : Searching for implicit temporal query results. *In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 2081–2083, New York, NY, USA. ACM.

Cantera, J. M., Arias, M., Cabrero, J., Garcia, G., Zubizarreta, A., Vegas, J. et de la Fuente, P. (2008). Mymose : Next generation search engine for mobile users. *In the 3rd edition of the Future of Web Search Workshop*.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F. et Li, H. (2007). Learning to rank : from pairwise approach to listwise approach. *In Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 129–136, New York, NY, USA. ACM.

Carterette, B., Kumar, N., Rao, A. et Zhu, D. (2011). Simple rank-based filtering for microblog retrieval : Implications for evaluation and test collections. *In Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011)*. National Institute of Standards and Technology (NIST).

Chang, A. X. et Manning, C. D. (2012). Sutime : A library for recognizing and normalizing time expressions. *In LREC*.

Chang, Y., Dong, A., Kolari, P., Zhang, R., Inagaki, Y., Diaz, F., Zha, H. et Liu, Y. (2013). Improving recency ranking using twitter data. *ACM Trans. Intell. Syst. Technol.*, pages 4 :1–4 :24.

Chapelle, O., Chang, Y. et Liu, T.-Y. (2011). Future directions in learning to rank. *In Yahoo ! Learning to Rank Challenge*, volume 14 de *JMLR Proceedings*, pages 91–100. JMLR.org.

Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E. et Yu, Y. (2012). Collaborative personalized tweet recommendation. *In Proceedings of the 35th international* ACM SIGIR *conference on Research and development in information retrieval*, SIGIR '12, pages 661–670, New York, NY, USA. ACM.

Chen, S. F. et Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. *In Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cheng, F., Zhang, X., He, B., Luo, T. et Wang, W. (2013). A survey of learning to rank for real-time twitter search. *In Proceedings of the 2012 International Conference on Pervasive Computing and the Networked World*, ICPCA/SWS'12, pages 150–164, Berlin, Heidelberg. Springer-Verlag.

Cheverst, K., Davies, N., Mitchell, K., Friday, A. et Efstratiou, C. (2000). Developing a context-aware electronic tourist guide : some issues and experiences. *In Proceedings of the* SIGCHI *conference on Human Factors in Computing Systems*, pages 17–24, New York, NY, USA. ACM.

Chisini, O. (1929). Sul concetto di media. (italian). *Periodico di Matematiche*, 9(4):106–116.

Choi, J. et Croft, W. B. (2012). Temporal models for microblogs. *In Proceedings of the 21st ACM International Conference on Information*

and Knowledge Management, CIKM '12, pages 2491–2494, New York, NY, USA. ACM.

Choquet, G. (1953). Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295.

Church, K. et Smyth, B. (2008). Who, what, where & when : a new approach to mobile search. *In Proceedings of the 2008 International Conference on Intelligent User Interfaces*, pages 309–312. ACM.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris.

Cong, G., Jensen, C. S. et Wu, D. (2009). Efficient retrieval of the top-k most relevant spatial web objects. *Proc. VLDB Endow.*, 2:337–348.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37.

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100.

Cormack, G. V., Clarke, C. L. A. et Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 758–759, New York, NY, USA. ACM.

Cosijn, E. et Ingwersen, P. (2000). Dimensions of relevance. *Information Processing and Management*, 36(4):533–550.

Costa, M., Couto, F. et Silva, M. (2014). Learning temporal-dependent ranking models. *In Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 757–766, New York, NY, USA. ACM.

Craswell, N. et Hawking, D. (2004). Overview of the TREC-2004 web track. *In Proceedings of the 13th* T*ext* RE*trieval* C*onference (*TREC *2004)*. National Institute of Standards and Technology (NIST).

Craswell, N., Robertson, S., Zaragoza, H. et Taylor, M. (2005). Relevance weighting for query independent evidence. *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 416–423, New York, NY, USA. ACM.

Craveiro, O., Macedo, J. et Madeira, H. (2009). Use of co-occurrences for temporal expressions annotation. *In Proceedings of the 16th International Symposium on String Processing and Information Retrieval*, SPIRE '09, pages 156–164, Berlin, Heidelberg. Springer-Verlag.

Cuadra, C. et Katter, R. (1967). Experimental study of relevance judgement. final report, OH : Case Western Reserve University, School of Library Science, Center for Documentation and Communication Research.

Cunningham, H., Maynard, D., Bontcheva, K. et Tablan, V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

da Costa Pereira, C., Dragoni, M. et Pasi, G. (2009). Multidimensional relevance : A new aggregation criterion. *In Proceedings of the 31st European Conference on Advances in Information Retrieval*, pages 264–275.

da Costa Pereira, C., Dragoni, M. et Pasi, G. (2012). Multidimensional relevance : Prioritized aggregation in a personalized information retrieval setting. *Information Processing and Management*, 48(2):340–357.

Dakka, W., Gravano, L. et Ipeirotis, P. G. (2012). Answering general time-sensitive queries. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):220–235.

Damak, F., Jabeur, L. B., Cabanac, G., Pinel-Sauvagnat, K., Tamine, L. et Boughanem, M. (2011). IRIT at TREC microblog 2011. *In Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011)*. National Institute of Standards and Technology (NIST).

Damak, F., Pinel-Sauvagnat, K., Boughanem, M. et Cabanac, G. (2013). Effectiveness of state-of-the-art features for microblog search. *In Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 914–919, New York, NY, USA. ACM.

Daoud, M. et Huang, J. X. (2013). Modeling geographic, temporal, and proximity contexts for improving geotemporal search. *Journal of the American Society for Information Science*, 64(1):190–212.

Daoud, M., Tamine, L. et Boughanem, M. (2010). A personalized graph-based document ranking model using a semantic user profile. *In Proceedings of the 18th international conference on User Modeling, Adaptation, and Personalization*, UMAP'10, pages 171–182, Berlin, Heidelberg.

Daoud, M., Tamine, L. et Mohand, B. (2011). A personalized search using a semantic distance measure in a graph-based ranking model. *Journal of Information Science (JIS)*, 37(6):614–636.

Dean-Hall, A., Clarke, C., Kamps, J., Thomas, P., Simone, N. et Voorhes, E. (2013). Overview of the trec 2013 contextual suggestion track. *In Text REtrieval Conference (TREC)*. National Institute of Standards and Technology (NIST).

Derczynski, L., Strötgen, J., Campos, R. et Alonso, O. (2015). Time and information retrieval : Introduction to the special issue. page 1–5.

Diaz, F. (2009). Integration of news content into web results. *In Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 182–191, New York, NY, USA. ACM.

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z. et Zha, H. (2010). Time is of the essence : Improving recency ranking using twitter data. *In Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 331–340, New York, NY, USA. ACM.

Duan, Y., Jiang, L., Qin, T., Zhou, M. et Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 295–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dubois, D. et Prade, H. (1996). Semantics of quotient operators in fuzzy relational databases. *Fuzzy Sets and Systems*, 78:89–93.

Dwork, C., Kumar, R., Naor, M. et Sivakumar, D. (2001). Rank aggregation methods for the web. *In Proceedings of the 10th International Conference on World Wide Web*, pages 613–622, New York, NY, USA. ACM.

Efron, M. (2010). Linear time series models for term weighting in information retrieval. *Journal of the Association for Information Science and Technology*, 61(7):1299–1312.

Efron, M. et Golovchinsky, G. (2011). Estimation methods for ranking recent information. *In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 495–504, New York, NY, USA. ACM.

Efron, M., Lin, J., He, J. et de Vries, A. P. (2014). Temporal feedback for tweet search with non-parametric density estimation. *In The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 33–42. ACM.

Eickhoff, C., de Vries, A. P. et Collins-Thompson, K. (2013a). Copulas for information retrieval. *In Proceedings of the 36th annual International* ACM SIGIR *Conference on Research and Development in Information Retrieval*, Dublin, Ireland. ACM.

Eickhoff, C., de Vries, A. P. et Collins-Thompson, K. (2013b). Copulas for information retrieval. *In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 663–672, New York, NY, USA. ACM.

Fagin, R., Kumar, R. et Sivakumar, D. (2003). Comparing top k lists. *In Proceedings of the fourteenth annual* ACM-SIAM *symposium on Discrete algorithms*, pages 28–36, Philadelphia, PA, USA.

Farah, M. et Vanderpooten, D. (2006). A multiple criteria approach for information retrieval. *In Proceedings International Symposium on String Processing and Information Retrieval (SPIRE 2006)*, Lecture Notes in Computer Science, pages 242–254. Springer Berlin Heidelberg.

Farah, M. et Vanderpooten, D. (2007). An outranking approach for rank aggregation in information retrieval. *In Proceedings of the 30th annual International* ACM SIGIR *Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 591–598, New York, NY, USA. ACM.

Farah, M. et Vanderpooten, D. (2008). An outranking approach for information retrieval. *Information Retrieval*, 11(4):315–334.

Filannino, M., Brown, G. et Nenadic, G. (2013). Mantime : Temporal expression identification and normalization in the tempeval-3 challenge.

Fox, E. A. et Shaw, J. A. (1993). *Combination of Multiple Searches*, pages 243–252. National Institute for Standards and Technology.

Frank, J. R., Kleiman-Weiner, M., Roberts, D. A., Niu, F., Zhang, C., Re, C. et Soboroff, I. (2012). Building an Entity-Centric Stream Filtering Test Collection for TREC 2012. *In Proceedings of the Text REtrieval Conference (TREC)*. NIST.

Gauch, S., Chaffee, J. et Pretschner, A. (2003). Ontology-based personalized search and browsing. *Web Intelli. and Agent Sys.*, 1(3-4):219–234.

Gerani, S., Zhai, C. et Crestani, F. (2012). Score transformation in linear combination for multi-criteria relevance ranking. *In Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pages 256–267, Berlin, Heidelberg. Springer-Verlag.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. et Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014.

Göker, A. et Myrhaug, H. (2008). Evaluation of a mobile information system in context. *Inf. Process. Manage.*, 44(1):39–65.

Grabisch, M. (1995). Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, 69(3):279–298.

Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89(3):445–456.

Grabisch, M., Kojadinovic, I. et Meyer, P. (2008). A review of methods for capacity identification in choquet integral based multi-attribute utility theory : Applications of the kappalab R package. *European Journal of Operational Research*, 186(2):766–785.

Grabisch, M. et Labreuche, C. (2010). A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *Annals OR*, 175(1):247–286.

Grabisch, M., Murofushi, T., Sugeno, M. et Kacprzyk, J. (2000). *Fuzzy Measures and Integrals. Theory and Applications.* Physica Verlag, Berlin.

Grabisch, M. et Nicolas, J.-M. (1994). Classification by fuzzy integral : performance and tests. *Fuzzy Sets Syst.*, 65(2-3):255–271.

Greengrass, E. (2000). Information retrieval : A survey.

Harper, S. et Chen, A. Q. (2012). Web accessibility guidelines : A lesson from the evolving web. *World Wide Web*, 15:1–28.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9):602–615.

Hattori, S., Tezuka, T. et Tanaka, K. (2007). Context-aware query refinement for mobile web search. *In Proceedings of the 2007 International Symposium on Applications and the Internet Workshops*, Washington, DC, USA. IEEE Computer Society.

Hwang, C.-L. et Yoon, K. (1981). *Multiple Attribute Decision Making. Methods and Applications : a State-of-the-Art Survey*. Springer-Verlag.

Jabeur, L. B., Tamine, L. et Boughanem, M. (2012). Uprising microblogs : A bayesian network retrieval model for tweet search. *In Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 943–948, New York, NY, USA. ACM.

Jankowski, P. (1995). Integrating geographical information systems and multiple criteria decision-making methods. *International Journal of Geographical Information Systems*, 9(3):251–273.

Jelinek, F. et Mercer, R. L. (1980). Interpolated estimation of markov source parameters from sparse data. *In In Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, The Netherlands : North-Holland.

Joachims, T. (2006). Training linear svms in linear time. *In Proceedings of the 12th* ACM SIGKDD *international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226, New York, NY, USA. ACM.

Joho, H., Jatowt, A., Blanco, R., Naka, H. et Yamamoto, S. (2014). Overview of ntcir-11 temporal information access (temporalia) task. *In Proceedings of the NTCIR-11 Conference*.

Joho, H., Jatowt, A. et Roi, B. (2013). A survey of temporal web search experience. *In Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 1101–1108, Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Joho, H. et Jose, J. M. (2008). Effectiveness of additional representations for the search result presentation on the web. *Information Processing & Management*, 44(1):226 – 241.

Jones, R. et Diaz, F. (2007). Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3).

Kanhabua, N. et Nørvåg, K. (2010). Determining time of queries for re-ranking search results. *In Research and Advanced Technology for Digital Libraries, 14th European Conference*, volume 6273, pages 261–272. Springer.

Kanhabua, N. et Nørvåg, K. (2011). A comparison of time-aware ranking methods. *In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1257–1258, New York, NY, USA. ACM.

Karkali, M., Rousseau, F., Ntoulas, A. et Vazirgiannis, M. (2014). Using temporal IDF for efficient novelty detection in text streams. *CoRR*, abs/1401.1456.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1):81–93.

Kim, H. D., Nikitin, D., Zhai, C., Castellanos, M. et Hsu, M. (2013). Information retrieval with time series query. *In Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages 14 :56–14 :63, New York, NY, USA. ACM.

Kishida, K. (2010). Vocabulary-based re-ranking for geographic and temporal searching at NTCIR geotime task. *In Proceedings of the 8th* NTCIR *Workshop Meeting on Evaluation of Information Access Technologies*, pages 181–184.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 91–101, New York, NY, USA. ACM.

Kolmogorov, A. N. (1930). On mean values. *Rend. Accad. dei Lincei*, 12(4):388–391.

Kulkarni, A., Teevan, J., Svore, K. M. et Dumais, S. T. (2011). Understanding temporal query dynamics. *In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 167–176, New York, NY, USA. ACM.

Larkey, L. S., Connell, M. E. et Callan, J. (2000). Collection selection and results merging with topically organized u.s. patents and TREC data. *In Proceedings of the ninth international conference on Information and*

knowledge management, CIKM '00, pages 282–289, New York, NY, USA. ACM.

Lee, J. H. (1997). Analyses of multiple evidence combination. *In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 267–276, New York, NY, USA. ACM.

Leung, C. W.-k., Chan, S. C.-f. et Chung, F.-l. (2006). A collaborative filtering framework based on fuzzy association rules and multiple-level similarity. *Knowl. Inf. Syst.*, 10(3):357–381.

Li, H. (2011). *Learning to Rank for Information Retrieval and Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Li, X. et Croft, W. B. (2003). Time-based language models. *In Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 469–475, New York, NY, USA. ACM.

Lin, J., Efron, M., Wang, Y. et Sherman, G. (2014a). Overview of the trec-2014 microblog track. *In Proceedings of the Text REtrieval Conference (TREC)*. NIST.

Lin, S., Jin, P., Zhao, X. et Yue, L. (2012). Tase : A time-aware search engine. *In Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2713–2715, New York, NY, USA. ACM.

Lin, S., Jin, P., Zhao, X. et Yue, L. (2014b). Exploiting temporal information in web search. *Expert Syst. Appl.*, 41(2):331–341.

Liu, F., Yu, C. et Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Trans. on Knowl. and Data Eng.*, 16(1):28–40.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

M., M., P., T., R., B., J., A., P., M. et Zaragoza., H. (2010). Searching through time in the new york times. *In Proceedings of the HCIR'10 Workshop*, page 41–44.

Ma, Z., Pant, G. et Sheng, O. R. L. (2007). Interest-based personalized search. *ACM Trans. Inf. Syst.*, 25(1).

Macdonald, C., Santos, R. et Ounis, I. (2013). The whens and hows of learning to rank for web search. *Information Retrieval*.

MacKay, D. J. et Peto, L. C. B. (1994). A hierarchical dirichlet language model. *Natural Language Engineering*, 1:1–19.

Mackay, D. J. C. et Peto, L. C. B. (1995). A hierarchical dirichlet language model. *Natural Language Engineering*, 1(3):289–308.

Manica, E., Dorneles, C. F. et Renata Galante, R. (2012). Handling temporal information in web search engines. *SIGMOD Rec.*, 41(3):15–23.

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. Chapman & Hall.

Marichal, J.-L. (1998). *Aggregation Operators for Multicriteria Decision Aid*. Thèse de doctorat, Institute of Mathematics, University of Liège, Liège, Belgium.

Marichal, J.-L. (2000). An axiomatic approach of the discrete choquet integral as a tool to aggregate interacting criteria. *Fuzzy Systems, IEEE Transactions on*, 8(6):800–807.

Marichal, J.-L. (2002). Aggregation operators. chapitre Aggregation of interacting criteria by means of the discrete Choquet integral, pages 224–244. Physica-Verlag GmbH, Heidelberg, Germany, Germany.

Maron, M. E. et Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244.

Massoudi, K., Tsagkias, M., de Rijke, M. et Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. *In Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 362–367, Berlin, Heidelberg. Springer-Verlag.

Mata, F. et Claramunt, C. (2011). Geost : geographic, thematic and temporal information retrieval from heterogeneous web data sources. *In Proceedings of the 10th international conference on Web and wireless geographical information systems*, pages 5–20, Berlin, Heidelberg. Springer-Verlag.

Mathews, L. K. et Kanmani, S. D. (2012). A survey on temporal information retrieval systems. *International Journal of Computer Applications*, 58(4):24–28.

Menger, K. (1942). Statistical metrics. *In Proceedings of the National Academy of Sciences of the United States of America*, 28(12):535–537.

Metzler, D. (2007). Automatic feature selection in the markov random field model for information retrieval. *In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007*, pages 253–262.

Metzler, D. et Cai, C. (2011). USC/ISI at TREC 2011 : Microblog track. *In Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011)*. National Institute of Standards and Technology (NIST).

Metzler, D., Jones, R., Peng, F. et Zhang, R. (2009). Improving search relevance for implicitly temporal queries. *In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 700–701, New York, NY, USA. ACM.

Miyanishi, T., Seki, K. et Uehara, K. (2012). Trec 2012 microblog track experiments at kobe university. *In Proceedings of the 21th* T*ext* RE*trieval* C*onference (*TREC *2012)*. National Institute of Standards and Technology (NIST).

Miyanishi, T., Seki, K. et Uehara, K. (2014). Time-aware latent concept expansion for microblog search. *In Proceedings of the Eighth International Conference on Weblogs and Social Media*, ICWSM'2014. The AAAI Press.

Mizzaro, S. (1998). How many relevances in information retrieval ? *Interacting with Computers*, 10(3):303–320.

Montgomery, D. C., Jennings, C. L. et Kulahci, M. (2008). *Introduction to Time Series Analysis and Forecasting*. Wiley Series in Probability and Statistics. Wiley, New York, NY.

Moulahi, B., Strötgen, J., Gertz, M. et Tamine, L. (2015a). Heideltoul : A baseline approach for cross-document event ordering. *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 825–829, Denver, Colorado. Association for Computational Linguistics.

Moulahi, B., Tamine, L. et Ben Yahia, S. (2013). L'intégrale de choquet discrète pour l'agrégation de pertinence multidimensionnelle. *In CORIA*, pages 399–414.

Moulahi, B., Tamine, L. et Ben Yahia, S. (2014a). IRIT at TREC 2014 Contextual Suggestion Track (regular paper). *In Text REtrieval Conference (TREC)*, page (on line), http ://www.nist.org. National Institute of Standards and Technology (NIST).

Moulahi, B., Tamine, L. et Ben Yahia, S. (2014b). Prise en compte des préférences des utilisateurs pour l'estimation de la pertinence multidimensionnelle d'un document. *In INFORSID*, pages 295–310.

Moulahi, B., Tamine, L. et Ben Yahia, S. (2014c). Toward a Personalized Approach for Combining Document Relevance Estimates (regular paper). *In Conference on User Modeling, Adaptation and Personalization (UMAP), Denmark, 07/07/2014-11/07/2014*, volume 8538 de *Lecture Notes in Computer Science*, pages 158–170. Springer.

Moulahi, B., Tamine, L. et Ben Yahia, S. (2015b). Leveraging Temporal Query-Term Dependency for Time-Aware Information Access (regular paper). *In IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE.

Moulahi, B., Tamine, L. et Ben Yahia, S. (2015c). When Time Meets Information Retrieval, Past Proposals, Current Plans, and Future Trends. *Journal of Information Science*.

Moulahi, B., Tamine, L. et Yahia, S. B. (2014d). iAggregator : Multidimensional relevance aggregation based on a fuzzy operator. *Journal of the Association for Information Science and Technology*, 65(10):2062–2083.

Murofushi, T. et Soneda, S. (1993). Techniques for reading fuzzy measures (iii) : Interaction index. *In Proceedings of the 9th Fuzzy Systems Symposium, Sapporo, Japan*, pages 693–696.

Nagmoti, R., Teredesai, A. et De Cock, M. (2010). Ranking approaches for microblog search. *In Proceedings of the 2010* IEEE/WIC/ACM *International Conference on Web Intelligence and Intelligent Agent Technology*, volume 01 de *WI-IAT '10*, pages 153–157, Washington, DC, USA. IEEE Computer Society.

Ng, A. Y., Jordan, M. I. et Weiss, Y. (2001). On spectral clustering : Analysis and an algorithm. *In ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press.

Nunes, S. (2007). Exploring Temporal Evidence in Web Information Retrieval. *In* MacFarlane, A., Azzopardi, L. et Ounis, I., éditeurs : *BCS*

IRSG Symposium Future Directions in Information Access (FDIA 2007), pages 44–50. BCS IRSG.

Nunes, S., Ribeiro, C. et David, G. (2008). Use of temporal expressions in web search. *In Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, pages 580–584, Berlin, Heidelberg. Springer-Verlag.

Nunes, S., Ribeiro, C. et David, G. (2011). Term weighting based on document revision history. *JASIST*, 62(12):2471–2478.

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C. et Ounis, I. (2012). Bieber no more : First story detection using twitter and wikipedia. *In SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012)*. ACM.

Ounis, T., Macdonald, J. et Soboroff, I. (2011). Overview of the TREC-2011 microblog track. *In Proceedings of the 20th* T*ext* RE*trieval* C*onference (*TREC *2011)*. National Institute of Standards and Technology (NIST).

Ounis, T., Macdonald, J. et Soboroff, I. (2012). Overview of the TREC-2012 microblog track. *In Proceedings of the 21th* T*ext* RE*trieval* C*onference (*TREC *2012)*. National Institute of Standards and Technology (NIST).

Perkiö, J., Buntine, W. et Tirri, H. (2005). A temporally adaptive content-based relevance ranking algorithm. *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 647–648, New York, NY, USA. ACM.

Pirkola, A. et Järvelin, K. (2001). Employing the resolution power of search keys. *JASIST*, 52(7):575–583.

Ponte, J. M. et Croft, W. B. (1998). A language modeling approach to information retrieval. *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

Porter, M. F. (1997). An algorithm for suffix stripping. *In* Sparck Jones, K. et Willett, P., éditeurs : *Readings in Information Retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pustejovsky, J. et Verhagen, M. (2009). Semeval-2010 task 13 : Evaluating events, time expressions, and temporal relations (tempeval-2). *In Proceedings of the Workshop on Semantic Evaluations : Recent Achievements and Future Directions*, DEW '09, pages 112–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radinsky, K., Svore, K., Dumais, S., Teevan, J., Bocharov, A. et Horvitz, E. (2012). Modeling and predicting behavioral dynamics on the web. *In Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 599–608, New York, NY, USA. ACM.

Radinsky, K., Svore, K. M., Dumais, S. T., Shokouhi, M., Teevan, J., Bocharov, A. et Horvitz, E. (2013). Behavioral dynamics on the web : Learning, modeling, and prediction. *ACM Trans. Inf. Syst.*, 31(3):16 :1–16 :37.

Rees, A. et Schultz, D. (1967). A field experiment approach to the study of relevance assessments in relation to document searching. Rapport technique, OH : Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University.

Ren, P., Chen, Z., Song, X., Li, B., Yang, H. et Ma, J. (2013). Understanding temporal intent of user query based on time-based query classification. *In Natural Language Processing and Chinese Computing*, volume 400, pages 334–345. Springer Berlin Heidelberg.

Renda, M. E. et Straccia, U. (2003). Web metasearch : rank vs. score based rank aggregation methods. *In Proceedings of the 2003 ACM Symposium on Applied Computing*, SAC '03, pages 841–846, New York, NY, USA. ACM.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd édition.

Riker, W. H. (1982). *Liberalism against populism*. Waveland Press.

Robertson, S. E. et Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3): 129–146.

Robertson, S. E. et Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *In Proceedings of the 17th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR '94, pages 232–241, New York, NY, USA. Springer-Verlag New York, Inc.

Robertson, S. E., Walker, S. et Hancock-Beaulieu, M. (1995). Large test collection experiments on an operational, interactive system : Okapi at TREC. *Inf. Process. Manage.*, 31(3):345–360.

Roy, B. (1991). The outranking approach and the foundations of electre methods. *Theory and Decision*, 31:49–73.

Roy, B. (2003). *Multicriteria Methodology for Decision Aiding*, volume 12 de *Nonconvex Optimization and Its Applications*. Springer US, 1 édition.

Roy, B., Vincke, P. et Brans, J. (1978). Aide à la décision multicritère. *Ricerca Operativa*, 8(5):11–45.

Salton, G. (1968). A comparison between manual and automatic indexing methods. Rapport technique, Ithaca, NY, USA.

Salton, G. (1971). *The SMART Retrieval System ; Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G. et McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Saracevic, T. (1976). Relevance : A review of the literature and a framework for thinking on the notion in information science. *In Advances in Librarianship*, pages 79–138. Academic Press.

Saracevic, T. (1996). Relevance reconsidered. *In Proceedings of the Second Conference on Conceptions of Library and Information Science (*CoLIS *2)*, pages 201–218.

Saracevic, T. (2007a). Relevance : A review of the literature and a framework for thinking on the notion in information science. part ii : nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):1915–1933.

Saracevic, T. (2007b). Relevance : A review of the literature and a framework for thinking on the notion in information science. part iii : Behavior and effects of relevance. *Journal of the American Society for Information Science*, 58(13):2126–2144.

Saracevic, T., Rothenberg, D. et Stephan, P. (1974). Study of infor-

mation utility. *In Proceedings of the American Society for information Science*, volume 11, pages 234–238.

Schamber, L. (1991). Users' criteria for evaluation in a multimedia environment. *In Proceedings of the 54th* ASIS *Annual Meeting*, 28:126–133.

Schilit, B. N., LaMarca, A., Borriello, G., Griswold, W. G., McDonald, D., Lazowska, E., Balachandran, A., Hong, J. et Iverson, V. (2003). Challenge : ubiquitous location-aware computing and the "place lab" initiative. *In Proceedings of the 1st* ACM *international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 29–35, New York, NY, USA. ACM.

Schweizer, B. et Sklar, A. (1960). Statistical metrics. *Pacific Journal of Mathematics*, 10(1):313–334.

Schweizer, B. et Sklar, A. (1983). *Probabilistic metric spaces*. North-Holland Series in Probability and Applied Mathematics. North-Holland Publishing Co., New York.

Shapley, L. S. (1953). A value for n-person games. *In* Kuhn, H. W. et Tucker, A. W., éditeurs : *Contributions to the Theory of Games*, volume 28 de *Annals of Mathematics Studies*, pages 307–317, Princeton. Princeton University Press.

Shokouhi, M. (2011). Detecting seasonal queries by time-series analysis. *In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1171–1172, New York, NY, USA. ACM.

Si, L. et Callan, J. (2002). Using sampled data and regression to merge search engine results. *In Proceedings of the 25th annual international* ACM SIGIR *conference on Research and development in information retrieval*, SIGIR '02, pages 19–26, New York, NY, USA. ACM.

Si, L. et Callan, J. (2003). A semisupervised learning method to merge search engine results. *ACM Trans. Inf. Syst.*, 21(4):457–491.

Sieg, A., Mobasher, B. et Burke, R. (2007). Web search personalization with ontological user profiles. *In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 525–534, New York, NY, USA. ACM.

Smith, M., Barash, V., Getoor, L. et Lauw, H. W. (2008). Leveraging social context for searching social media. *In Proceedings of the 2008* ACM

workshop on Search in social media, pages 91–94, New York, NY, USA. ACM.

Sokolan, P., Doherty, D., Duguay, C., Radcliffe, W. et Bourassa, V. (2015). Search result presentation. US Patent 8,973,128.

Song, F. et Croft, W. B. (1999). A general language model for information retrieval. *In In Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 279–280.

Steuer, R. E. (1986). *Multiple Criteria Optimization : Theory, Computation and Application.* John Wiley & Sons, New York.

Strötgen, J., Alonso, O. et Gertz, M. (2012). Identification of top relevant temporal expressions in documents. *In Proceedings of the 2Nd Temporal Web Analytics Workshop*, TempWeb '12, pages 33–40, New York, NY, USA. ACM.

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Inf. Process. Manage.*, 28(4):503–516.

Su, L. T. (1994). The relevance of recall and precision in user evaluation. *J. Am. Soc. Inf. Sci.*, 45(3):207–217.

Subasic, I. et Castillo, C. (2010). The effects of query bursts on web search. *In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1 de *WI-IAT '10*, pages 374–381, Washington, DC, USA. IEEE Computer Society.

Sugeno, M. (1974). *Theory of fuzzy integrals and its applications.* Thèse de doctorat, Tokyo Institute of Technology.

Sugeno, M. (1977). Fuzzy measures and fuzzy integrals : a survey. *In* Gupta, M., Saridis, G. et Gains, B., éditeurs : *Fuzzy automata and decision processes*, pages 89–102. North Holland, Amsterdam.

Swan, R. et Allan, J. (2000). Automatic generation of overview timelines. *In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 49–56, New York, NY, USA. ACM.

Taylor, A. R. (2012). User relevance criteria choices and the information search process. *Information Processing and Management*, 48(1):136–153.

Taylor, A. R., Cool, C., Belkin, N. J. et Amadio, W. J. (2007). Relationships between categories of relevance criteria and stage in task completion. *Information Processing and Management*, 43(4):1071–1084.

Taylor, R. S. (1986). *Value-added processes in information systems*. Ablex Pub. Corp.

UzZaman, N., Llorens, H., Allen, J. F., Derczynski, L., Verhagen, M. et Pustejovsky, J. (2012). Tempeval-3 : Evaluating events, time expressions, and temporal relations. *CoRR*, abs/1206.5333.

Vallet, D. et Castells, P. (2012). Personalized diversification of search results. *In Proceedings of the 35th annual International* ACM SIGIR *Conference on Research and Development in Information Retrieval*, pages 841–850. ACM.

Vansnick, J.-C. (1986). Mathematical programming multiple criteria decision making on the problem of weights in multiple criteria decision making (the noncompensatory approach). *European Journal of Operational Research*, 24(2):288 – 294.

Verhagen, M., Saurí, R., Caselli, T. et Pustejovsky, J. (2010). Semeval-2010 task 13 : Tempeval-2. *In Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vickery, B. C. (1959). Subject analysis for information retrieval. *In Proceedings of the International Conference on Scientific Information*, volume 2, pages 855–865.

Vlachos, M., Meek, C., Vagena, Z. et Gunopulos, D. (2004). Identifying similarities, periodicities and bursts for online search queries. *In Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pages 131–142, New York, NY, USA. ACM.

Vogt, C. C. et Cottrell, G. W. (1999). Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173.

Wang, P., Berry, M. W. et Yang, Y. (2003). Mining longitudinal web queries : Trends and patterns. *J. Am. Soc. Inf. Sci. Technol.*, 54(8):743–758.

Wang, S., Lu, K., Lu, X. et Wang, B. (2014). Query dependent time-sensitive ranking model for microblog search. *In Web Technologies and*

Applications, volume 8709 de *Lecture Notes in Computer Science*, pages 644–651. Springer International Publishing.

Wei, C.-P., Lee, Y.-H., Chiang, Y.-S., Chen, C.-T. et Yang, C. C. (2014). Exploiting temporal characteristics of features for effectively discovering event episodes from news corpora. *Journal of the Association for Information Science and Technology*, 65(3):621–634.

Wei, F., Li, W. et Liu, S. (2010). iRANK : A rank-learn-combine framework for unsupervised ensemble ranking. *Journal of the American Society for Information Science and Technology*, 61(6):1232–1243.

Wei, Z., Zhou, L., Li, B., Wong, K.-F., Gao, W. et Wong, K.-F. (2011). Exploring tweets normalization and query time sensitivity for twitter search. *In Proceedings of The Twentieth Text REtrieval Conference*.

Weinberger, K., Mohan, A. et Z., C. (2010). Tree ensembles and transfer learning. *In Proceedings of the Yahoo ! Learning to Rank Challenge Workshop*, WWW '10.

White, R. W., Ruthven, I. et Jose, J. M. (2005). A study of factors affecting the utility of implicit relevance feedback. *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 35–42, New York, NY, USA. ACM.

Wolfe, S. R. et Zhang, Y. (2010). Interaction and personalization of criteria in recommender systems. *In Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization*, UMAP'10, pages 183–194, Berlin, Heidelberg. Springer-Verlag.

Wu, Q., Burges, C. J., Svore, K. M. et Gao, J. (2010). Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3):254–270.

Xu, J. et Li, H. (2007). Adarank : A boosting algorithm for information retrieval. *In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 391–398, New York, NY, USA. ACM.

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions On Systems Man And Cybernetics*, 18(1):183–190.

Yau, S. S., Liu, H., Huang, D. et Yao, Y. (2003). Situation-aware personalized information retrieval for mobile internet. *In Proceedings of the 27th*

Annual International Conference on Computer Software and Applications, pages 638–, Washington, DC, USA. IEEE Computer Society.

Yu, C. T. et Salton, G. (1976). Precision weighting - an effective automatic indexing method. *J. ACM*, 23(1):76–88.

Yu, P. S., Li, X. et Liu, B. (2004). On the temporal dimension of search. *In Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, WWW Alt. '04, pages 448–449, New York, NY, USA. ACM.

Zhai, C. et Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214.

Zhang, C., Xu, W., Meng, F., Li, H., Wu, T. et Xu, L. (2013). The information extraction systems of PRIS at temporal summarization track. *In Text REtrieval Conference (*TREC*)*. NIST.

Zhang, G., Yang, Z. et Si, S. (2014). Ebjut at trec 2014 microblog track. *In Proceedings of the Text REtrieval Conference (TREC)*. NIST.

Zhao, L., Zeng, Y. et Zhong, N. (2011). A weighted multi-factor algorithm for microblog search. *In Proceedings of the 7th International Conference on Active Media Technology*, AMT'11, pages 153–161, Berlin, Heidelberg. Springer-Verlag.

Zhu, Y. et Shasha, D. (2003). Efficient elastic burst detection in data streams. *In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 336–345, New York, NY, USA. ACM.

Zhu, Y., Xue, Y., Guo, J., Lan, Y., Cheng, X. et Yu, X. (2012). Exploring and exploiting proximity statistic for information retrieval model. *In Information Retrieval Technology*, volume 7675 de *Lecture Notes in Computer Science*, pages 1–13. Springer Berlin Heidelberg.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.